# NAVAL RESEARCH LOGISTICS QUARTERLY

JUNE 1969
VOL. 16, NO. 2


OFFICE OF NAVAL RESEARCH

NAVSO P-1278

# NAVAL RESEARCH LOGISTICS QUARTERLY




The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.


The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $5.50 a year in the U.S. and Canada, $7.00 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.





Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, NAVEXOS P–35

# BAYES SEQUENTIAL DESIGN OF STOCK LEVELS [1]

S. Zacks

*The University of New Mexico*

**ABSTRACT**

Bayesian determination of optimal stock levels is studied for the case of Poisson distribution of the demand variable, and prior gamma distribution of the expected demand. Bayes sequential procedure is derived, assuming that stock level can be adjusted at the beginning of each period so that a shortage can be immediately replenished and an overstock can be corrected. The Bayes sequential procedure is more difficult to obtain if this assumption is removed. A dynamic programming method for obtaining the general Bayes sequential procedure is outlined. Finally, an empiric Bayes estimation procedure of the optimal Bayesian stock level is presented.

## 1. INTRODUCTION

This paper discusses the Bayes sequential determination of optimal stock levels. Studies of Bayes sequential determination of stock levels have been published by Karlin [2], Scarf [5] and others, and are reviewed by Veinott [6]. The general solution is, as shown in Section 6, a complicated one. In Sections 3 and 4, we give a more simple solution, which results from an assumption concerning the adjustability of the stock level at the end of each period. We start by exhibiting the Bayesian solution to the case where the planning horizon spans over one period, and then we extend the result to the case of $N \geq 1$ periods. The last section, Section 7, is devoted to an outline of an empiric Bayes approach, when the prior distribution is not exactly known.

The following inventory model is presumed:

1. The demand variables in each period are independent and identically distributed like a Poisson random variable, with mean $\theta$.

2. The prior distribution of $\theta$ is assumed to be a gamma,

$$\mathcal{G}\left(\frac{1}{\tau}, \nu\right), \ 0 < \tau < \infty, \ 0 < \nu < \infty,$$

with mean $\nu\tau$.

3. There is no lead time, and orders are replenished instantaneously at the beginning of each period.

4. There are no backlogs. The stock level at the start of each period is designed so as to avoid losses during that period. Shortages or overstockings can be adjusted and corrected at the beginning of each period. (In Section 6, a solution is provided which does not assume overstocking adjustment.)

Although the present model may seem to be oversimplified, it is sufficient for sub-inventory systems (e.g., the submarine) which can be adjusted periodically at the base inventory (e.g., the tender boat). It is not a satisfactory or complete model for a more complicated inventory system (e.g., the warehouse).

[1] This paper is a revised version of the Technical Memorandum, Serial TM-14424, Logistics Research Project, The George Washington University.

We assume the common convex piece-wise linear cost function

$$L(x;k)=\begin{cases}c(k-x), & x \leq k \\ p(x-k), & x > k,\end{cases} \tag{1.1}$$

where $0 < c, p < \infty$; $k$ designates the stock level at the beginning of a period, $x$ denotes the demand value. $c$[\$] is the carrying cost of one unit per one time period; $p$[\$] is the penalty cost per unit in shortage of demand.

If $(k_1, \ldots, k_N)$ denote the stock levels assigned to the $N$ periods under consideration, the objective is to determine these stock levels so as to minimize the total expected prior loss, namely:

$$R(\tau, \nu; k_1, \ldots, k_N) = E\left\{\sum_{i=1}^{N} L(X_i; k_i)\right\}. \tag{1.2}$$

In the sequential procedure, the value of $k_i$ $(i=2, \ldots, N)$ is determined as a function of $(\tau, \nu X_1, \ldots X_{i-1})$; where $X_i (i=1, 2, \ldots)$ is the demand values at the $i$-th period. $k_1$ is a function of $(\tau, \omega)$ only.

As shown in Section 4, for the above simple model, the optimal stock level for the $n$th period, given the demand values $(X_1, \ldots, X_{n-1})$, is the $p/(c+p)$th fractile of the posterior marginal distribution of $X_n$. This posterior marginal distribution is the mixture of the Poisson distribution, $\mathscr{P}(\theta)$, with respect to the posterior distribution of $\theta$, given $(X_1, \ldots, X_{n-1})$.

As shown in Section 2, the posterior marginal distribution of $X_n$ is a negative binomial distribution. In Section 3 we give a method for determining the fractiles of a negative binomial distribution from tables of the percentage points of the beta distribution function [1], or from tables of the incomplete beta function [3].In case the parameter $\nu$ of the prior gamma distribution of $\theta$ is an integer one can use existing tables of the binomial distribution. A numerical example is given in Section 3.

## 2. PRIOR, POSTERIOR, AND MARGINAL DISTRIBUTIONS.

In the Bayesian framework, we consider the parameter $\theta$ of the Poisson distribution as a random variable. In this paper, we restrict attention to cases where the prior distribution of $\theta$ is a member of the family of gamma distributions, i.e.,

$$\mathfrak{L}(\theta) = \mathscr{G}\left(\frac{1}{\tau}, \nu\right), \quad 0 < \tau < \infty, \quad 0 < \nu < \infty, \tag{2.1}$$

where $1/\tau$ is the scale parameter, and the prior expectation of $\theta$ is $\nu\tau$.

It is well known that, if $X_1, X_2, \ldots, X_n$ are independent random variables having an identical Poisson distribution $\mathscr{P}(\theta)$, then

$$S_n = \sum_{i=1}^{n} X_i$$

is a complete-sufficient statistic, having a Poisson distribution $\mathscr{P}(n\theta)$. Furthermore, it is easy to verify that, if $\mathscr{G}(1/\tau, \nu)$ is the prior distribution of $\theta$, then the posterior distribution law of $\theta$ given $S_n$ is again gamma, namely:

$$\mathfrak{L}(\theta | S_n) = \mathscr{G}\left(\frac{1}{\tau} + n, S_n + \nu\right). \tag{2.2}$$

Another distribution of special importance is the posterior marginal distribution of $X_{n+1}$ given $S_n$. This is the mixture of the Poisson distribution $\mathscr{P}(\theta)$ with respect to the posterior distribution of $\theta$,

$\mathcal{G}(1/\tau+n, S_n+\nu)$. It is a straightforward matter to verify that the posterior marginal distribution of $X_{n+1}$, given $S_n$, is the negative binomial distribution, with the density function

$$g(x|\psi_{n+1}, \nu_{n+1}) = \frac{\Gamma(x+\nu_{n+1})}{\Gamma(\nu_{n+1})\Gamma(x+1)} \psi_{n+1}^{x} (1-\psi_{n+1})^{\nu_{n+1}}, \tag{2.3}$$

for x = 0, 1, 2, . . ., where

$$\nu_{n+1} = S_n + \nu, \qquad n = 0, 1, \ldots$$

and

$$\psi_{n+1} = \tau/(1+(n+1)\tau), \qquad n+0, 1, \ldots ..$$

$S_0$ is defined as 0. It should be noted that, whenever $\nu_n = 1$ the above negative-binomial distribution reduces to a geometric distribution; and if $\nu_n$ is an integer then the negative-binomial distribution is a Pascal distribution.

## 3. THE BAYES DETERMINATION OF INVENTORY LEVEL FOR N = 1

Suppose that we have to determine the inventory level $k$, only for one period, i.e., $N = 1$. Suppose also that we have not observed $X$ as yet. The optimal stock level $k_1^0$ is the least integer $k$, minimizing the prior risk, namely, the expected loss under the prior marginal $g(x|\psi_1, \nu_1)$. The prior risk function, for $N = 1$, given $k$, is thus:

$$R_1(k; \psi_1, \nu_1) = \sum_{x=0}^{\infty} L(k; x) g(x/\psi_1, \nu_1), \tag{3.1}$$

where $\psi_1 = \tau/(1+\tau)$, $\nu_1 = \nu$.

Consider the expected loss under any distribution of $X$, say $P(x)$. We have,

$$R(k; P) = \int_{\{x \le k\}} c(k-x)dP(x) + \int_{\{x>k\}} p(x-k)dP(x). \tag{3.2}$$

Differentiating (3.2) formally with respect to $k$, and equating the partial derivative to zero, we obtain the well known result (see Veinott [6]) that the optimal level $k_1^0$ is given approximately by the $p/(c+p)$th fractile of $P(x)$. For the present case, we have to determine the $p/(c+p)$th fractile of the negative-binomial distribution, with parameters $(\psi_1, \nu_1)$. Generally, let $G_\gamma(\psi, \nu)$ designate the $\gamma$-th fractile, $0 < \gamma < 1$, of the negative-binomial distribution having parameters $\psi$ and $\nu$ (replace $\psi_{n+1}$ by $\psi$ and $\nu_n$ by $\nu$ in (2.3)). Thus, the optimal Bayes stock level, $k_1^0$, for $N = 1$, is

$$k_1^0 \cong G_{\frac{p}{c+p}}(\psi_1, \nu_1). \tag{3.3}$$

We show now how $k_1^0$ can be determined from tables of the incomplete beta function ratio or, when $\nu_1$ is an integer, by tables of the binomial distribution. Let $G(k|\psi, \nu)$ designate the negative binomial distribution with parameters $(\psi, \nu)$, i.e.,

$$G(k|\psi, \nu) = \sum_{x=0}^{k} \frac{\Gamma(x+\nu)}{\Gamma(\nu)\Gamma(x+1)} \psi^x (1-\psi)^{\nu}. \tag{3.4}$$

The $\gamma$-th fractile ($0<\gamma<1$) of the negative binomial distribution is a non-negative integer satisfying:

$$G_\gamma(\psi, \nu) = \min \{k: k \geq 0 \text{ and } G(k|\psi, \nu) \geq \gamma\}. \tag{3.5}$$

As proven in the appendix,[2] the value of $G(k|\psi, \nu)$ can be expressed by means of the incomplete beta function ratio

$$I_\theta(p, q) = \frac{1}{B(p, q)} \int_0^\theta \mu^{p-1}(1-\mu)^{q-1} d\mu, \quad 0 < p, q < \infty,$$

by the following formula:

$$\begin{aligned} G(k|\psi, \nu) &= \sum_{x=0}^{k} \frac{\Gamma(x+\nu)}{\Gamma(\nu)\Gamma(x+1)} \psi^x (1-\psi)^\nu \\ &= I_\theta(\nu, k+1), \end{aligned} \tag{3.6}$$

where $\theta = 1-\psi$. $\nu$ is not necessarily an integer, but $k$ is a non-negative integer. Hence, if tables of the incomplete beta-function ratio $I_\theta(v, k+1)$ are unavailable, one can use the formula (see Harter [1]).

$$I_\theta(\nu, k+1) = \frac{\Gamma(\nu+k+1)}{\Gamma(\nu)\Gamma(k+1)} \theta\nu \sum_{i=0}^{k} (-1)^i \binom{k}{i} \frac{\theta^i}{\nu+i}. \tag{3.7}$$

If $\nu$ is an integer, we can utilize the relationship between the incomplete beta function ratio and the binomial distribution to express (3.6) in the form:

$$G(k|\psi, v) = \sum_{x=\nu}^{\nu+k} \binom{\nu+k}{x} (1-\psi)^x \quad \psi^{\nu+k-x}. \tag{3.8}$$

For large values of $\nu$, we can approximate (3.7) by a normal distribution to obtain:

$$G(k|\psi, \nu) \approx 1 - \Phi\left(\frac{\nu - (\nu+k)(1-\psi)}{\sqrt{(v+k)\psi(1-\psi)}}\right), \text{ as } \nu \to \infty. \tag{3.9}$$

The function $\Phi(\cdot)$ in (3.9) is the standard normal integral. Whatever formula or tables we use to determine the incomplete beta function ratio $I_\theta(\nu, k+1)$, we can write the general formula of the $G_\gamma(\psi, \nu)$ as:

$$G_\gamma(\psi, \nu) = \text{least non-negative integer } k, \tag{3.10}$$

such that:

$$I_{1-\psi}(\nu, k+1) \geq \gamma.$$

For example, consider the case of $\tau = 1$, $\nu = 3$. Thus, $\psi = \tau/1(1+\tau) = 1/2$. Suppose also that $c$ and $p$ are such that $\gamma = p/(c+p) = 0.6$. The 0.6-fractile of the negative-binomial distribution with parameters $\psi = 1/2$ and $\nu = 3$, $G_{0.6}(1/2, 3)$ is the least non-negative integer $k$, such that $I_{0.5}(3, k+1) \geq 0.6$. Using the relationship

$$I_\theta(a, b) = 1 - I_{1-\theta}(b, a)$$

we obtain that $G_{0.6}(1/2, 3)$ is the least non-negative interger $k$, such that $I_{0.5}(k+1, 3) \leq 0.4$. Let us denote by $X(0.4; k+1, 3)$ the 0.4th fractile of the beta $(k+1, 3)$ distribution. Thus, since $I_{0.5}(k+1, 3) \leq 0.4$ implies that $X(0.4; k+1, 3) \geq 0.5$ we obtain that $G_{0.6}(1/2, 3) =$ least non-negative integer $k$,

[2] The relationship (3.6) is proven also in Raiffa and Schlaifer [4], p. 237.

such that $X(0.4;\ k+1,\ 3) \geq 0.5$. From the tables of Harter [1] we compile the following values:

| $k$ | $X(0.4;\ k+1,\ 3)$ |
|---|---|
| 0 | 0.1565 |
| 1 | 0.3292 |
| 2 | 0.4462 |
| 3 | 0.5292 |

Hence, $G_{0.6}\ (1/2,\ 3) = 3$.

Harter's tables can be used, as illustrated here for integer values of $\nu = 1(1)40$ and for $\gamma = 0.5$, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.995 and 0.999. The general rule that should be used is: $G_\gamma(\psi, \nu) =$ least integer $k$ such that $X(1-\gamma;\ k+1,\ \nu) \geq \psi$. In Table 1, we present the $k^\circ\ (\psi, \nu)$ values for several integer values of $\nu$ and five values of $\psi$, with $\gamma = p/(c+p) = 0.6$.

TABLE 1. *Values of* $k^\circ(\psi, \nu)$ *for* $p/(c+p) = 0.6$

| $\nu$ \ $\psi$ | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 2 | 8 |
| 2 | 0 | 1 | 2 | 5 | 18 |
| 3 | 0 | 1 | 3 | 7 | 28 |
| 4 | 0 | 2 | 4 | 10 | 38 |
| 5 | 1 | 2 | 5 | 12 | *51 |
| 10 | 1 | 5 | 11 | 25 | *98 |
| 20 | 2 | 9 | 21 | *48 | *191 |
| 30 | 4 | 14 | 31 | *71 | *284 |

All the values in Table 1 which are not starred were compiled from Harter's Tables [1]. The starred values were obtained by a normal approximation.

## 4. THE GENERAL CASE OF N PERIODS.

In this section, we treat the general case, where the planning horizon extends over $N$ periods, $N \geq 1$. We set up the solution in terms of the dynamic programming backward induction method, and show that for the inventory model described in the introduction the optimal stock level for the beginning of the $n$-th period ($n = 1, 2, \ldots$) is the $p/(c+p)$th fractile of the negative binomial distribution, with parameters $\psi_n = \tau/(1+n\tau)$ and $\nu_n = \nu + S_{n-1}$. We also show how to determine the Bayes risk of the optimal procedure.

Since $(S_0, S_1, S_2, \ldots, S_{n-1})$ is a set of sufficient statistics for the truncated (after $N$ periods) sequential model, the optimal stock level for the beginning of the $n$-th month ($n = 1, \ldots, N$) is a function $k_n^0(\psi_n, \nu_n)$. Assumption 4 of the model presented in Section 1 states that $k_n^0(\psi_n, \nu_n)$ can be smaller than $k_{n-1}^0(\psi_{n-1}, \nu_{n-1}) - X_{n-1}$, if the last expression is positive, for every $n = 2, 3, \ldots, N$. This assumption leads to a considerable simplification of the dynamic programming recursive relations, and hence to a relatively simple solution. In Section 6, we discuss the solution under the restriction of $k_n^0(\psi_n, \nu_n) \geq k_{n-1}^0(\psi_{n-1}, \nu_{n-1}) - X_{n-1}$.

Consider first the optimal stock level for the $N$th period. Whatever was our stocking policy during the first $(n-1)$ periods, only one period is left at the planning horizon, and we could therefore apply the solution of the previous section, with $\psi_1$ and $\nu_1$ replaced by $\psi_N$ and $\nu_N$. Hence, the optimal stock level for the beginning of the $N$th period is approximately the $p/(c+p)$th fractile of the corresponding negative binomial distribution, that is:

$$k_N^0(\psi_N, \nu_N) \approx G_{\frac{p}{c+p}}(\psi_N, \nu_N). \tag{4.1}$$

This value can be determined by formula (3.10) in which we replace $\psi_1$ and $\nu_1$ by $\psi_N$ and $\nu_N$. Let $\rho o(\psi_N, \nu_N)$ designate the Bayes risk associated with the optimal stock level of the $N$th period. That is,

$$\rho_o(\psi_N, \nu_N) = \sum_{x=0}^{\infty} L(x, k_N^0(\psi_N, \nu_N)) g(x|\psi_N, \nu_N). \tag{4.2}$$

As shown in Appendix 2, for any $(\psi, \nu)$,

$$\rho_0(\psi, \nu) \cong (c+p) \frac{\Gamma(k^0(\psi, \nu)+\nu+1)}{\Gamma(\nu)\Gamma(k^0(\psi, \nu)+1)} \psi^{k^0(\psi,\nu)+1}(1-\psi)^{\nu-1}, \tag{4.3}$$

where $k^0(\psi, \nu)$ is the $p/(c+p)$th fractile of the negative binomial distribution, with parameters $(\psi, \nu)$. In the sequel, we shall identify $\rho_0(\psi, \nu)$ with the R.H.S. of (4.3).

We consider now the determination of one optimal stock level for the $(N-1)$st period. Given $(\psi_{N-1}; \nu_{N-1})$ the optimal stock level $k_{N-1}^0(\psi_{N-1}, \nu_{N-1})$ is a non-negative integer which minimizes the expected posterior risk of the last two periods, given that $k_N^0(\psi_N, \nu_N)$ is (4.1). Thus, we have to minimize the function:

$$R_1(k; \psi_{N-1}, \nu_{N-1}) = E_{x_{N-1}|\psi_{N-1}, \nu_{N-1}}\{L(X_{N-1}; k) + \rho_0(\psi_N, \nu_{N-1}+X_{N-1})\}. \tag{4.4}$$

$E_{X|\psi,\nu}\{\ \}$ designates the expectation of the function of $X$ in the brackets, with respect to the negative binomial distribution, with parameters $(\psi, \nu)$.

We define the Bayes risk for the last 2 months as

$$\rho_1(\psi_{N-1}, \nu_{N-1}) = \inf_{0 \le k < \infty} R_1(k; \psi_{N-1}, \nu_{N-1}). \tag{4.5}$$

Since the distribution $G(x|\psi_{N-1}, \nu_{N-1})$ does not depend on the stock level $k_{N-1}$, we obtain from (4.4) that

$$\rho_1(\psi_{N-1}, \nu_{N-1}) = \inf_{0 \le k \le \infty} E_{X_{N-1}|\psi_{N-1}, \nu_{N-1}}\{L(X_{N-1}; k)\} + E_{X_{N-1}|\psi_{N-1}, \nu_{N-1}}\{\rho_0(\psi_N, \nu_{N-1}+X_{N-1})\}. \tag{4.6}$$

The value of $k$ for which (4.6) is attained is the (Bayes) optimal stock level for the $(N-1)$st period. From (4.6) we immediately obtain that,

$$k_{N-1}^0(\psi_{N-1}, \nu_{N-1}) \cong G_{\frac{p}{c+p}}(\psi_{N-1}, \nu_{N-1}). \tag{4.7}$$

Furthermore, the Bayes risk (4.6) is

$$\rho_1(\psi_{N-1}, \nu_{N-1}) = \rho_0(\psi_{N-1}, \nu_{N-1}) + \sum_{x=0}^{\infty} \rho_0(\psi_N, \nu_{N-1}+x) g(x|\psi_{N-1}, \nu_{N-1}). \tag{4.8}$$

In general, by the backward induction principle of the dynamic programming method, we define the functions:

$$R_j(k;\psi_{N-j},\nu_{N-j})=E_{X_{N-j}|\psi_{N-j},\nu_{N-j}}\{L(X_{N-j};k)+\rho_{j-1}(\psi_{N-j+1},\nu_{N-j}+X_{N-j})\},\qquad j=1,2,\ldots,N \tag{4.9}$$

and

$$\rho_j(\psi_{N-j},\nu_{N-j})=\inf_{0\le k<\infty}R_j(k;\psi_{N-j},\nu_{N-j}),\qquad j=1,2,\ldots,N. \tag{4.10}$$

The same argument as before leads immediately to the conclusion that

$$k^0_{N-j}(\psi_{N-j},\nu_{N-j})\cong G_{\frac{p}{c+p}}(\psi_{N-j},\nu_{N-j}), \tag{4.11}$$

for all $j=0,1,\ldots,N-1$; and that

$$\rho_j(\psi_{n-j},\nu_{N-j})=\rho_0(\psi_{N-j},\nu_{N-j})+E_{X_{N-j}|\psi_{N-j},\nu_{N-j}}\{\rho_{j-1}(\psi_{N-j+1},\nu_{N-j}+X_{N-j})\},\qquad j=0,\ldots,N-1. \tag{4.12}$$

$\rho_{N-1}(\psi_1,\nu_1)$ is the Bayes risk associated with the prior gamma distribution $\mathcal{G}(1/\tau,\nu)$, and the optimal Bayes sequential rule that the stock level at the beginning of the $n$-th period ($n=1,2,\ldots,N$) is the $p/(c+p)$th fractile of the negative binomial distribution, the parameters of which are adjusted after each period to $(\psi_n,\nu_n)$. As noticed already, $\nu_n=\nu+S_{n-1}$ is the statistic summarizing the demand values in the previous periods. Repeated application of (4.12) yields

$$\rho_{N-1}(\psi_1,\nu_1)=\rho_0(\psi_1,\nu_1)+\sum_{j=2}^{N}E_{S_{j-1}|\psi_1,\nu_1}\{\rho_0(\psi_j,\nu_1+S_{j-1})\}. \tag{4.13}$$

Now, the distribution of $S_{j-1}$ given $\theta$ is Poisson with mean $(j-1)\theta$. Hence, the posterior marginal distribution of $S_{j-1}$ given $(\psi_1,\nu_1)$ is the negative binomial, with parameters $(\tilde{\tilde{\psi}}_{j-1},\nu_1)$ where

$$\tilde{\tilde{\psi}}_{j-1}=(j-1)\psi_{j-1},\qquad j=2,\ldots,N. \tag{4.14}$$

Hence, we obtain from (4.13) that

$$\rho_{N-1}(\psi_1,\nu_1)=\rho_0(\psi_1,\nu_1)+\sum_{j=2}^{N}\sum_{x=0}^{\infty}\rho_0(\psi_j,\nu_1+x)g(x|\tilde{\tilde{\psi}}_{j-1},\nu_1). \tag{4.15}$$

After the value of $X_1$ associated with the first period has been observed, it is illogical to measure the effectiveness of the procedure by $\rho_{N-1}(\psi_1,\nu_1)$. For this case, we calculate (if $N\ge 3$)

$$\rho_{N-2}(\psi_2,\nu_2)=\rho_0(\psi_2,\nu_2)+\sum_{j=3}^{N}\sum_{x=0}^{\infty}\rho_0(\psi_j,\nu_2+x)g(x|\tilde{\psi}_{j-2},\nu_2), \tag{4.16}$$

where $\nu_2=\nu_1+X_1$; $\tilde{\psi}_{j-2}=(j-2)\psi_{j-2}(j=3,\ldots,N)$.

Generally, the Bayes risk function relevant for characterizing the efficiency of the Sequential Bayes procedure at the beginning of the $n$-th month is

$$\rho_{N-n}(\psi_n,\nu_n)=\begin{cases}\rho_0(\psi_N,\nu_N)\\ \rho_0(\psi_n,\nu_n)+\sum_{j=n+1}^{N}\sum_{x=0}^{\infty}\rho_0(\psi_j,\nu_n+x)g(x|\tilde{\psi}_{j-n},\nu_n),\ \text{if } n=N,\end{cases} \tag{4.17}$$

in which $\tilde{\psi}_{j-n}=(j-n)\psi_{j-n}$, for $j=n+1,\ldots,N$, and $\nu_n=\nu+S_{n-1}$.

## 5. THE CASE OF INFINITE HORIZON.

As seen in the previous section, the optimal stock level $k_n^0(n=1, 2, \ldots)$ is a function of the parameters $\psi_n$ and $\nu_n$, which depend only on the prior information and the total previous demand. $k_n^0(\psi_n, \nu_n)$, however, is independent of the number of periods in the horizon $N$. Hence

$$k_n^0(\psi_n, \nu_n) \cong G_\gamma(\psi_n, \nu_n)$$

with $\gamma = p/(c+p)$ is the optimal stock level for arbitrarily large $N$ (infinite horizon).

Let $\rho_\infty(\psi, \nu)$ designate the Bayes risk associated with the optimal policy, when $(\psi, \nu)$ are the prior parameters. From (4.12) we obtain, due to the monotonicity of $\rho_{N-1}(\psi, \nu)$ as a function of $N$, the functional equation:

$$\rho_\infty(\psi, \nu) = \rho_0(\psi, \nu) + \sum_{x=0}^{\infty} g(x|\psi, \nu)\rho_\infty\left(\frac{\psi}{1+\psi}, \nu+x\right), \tag{5.1}$$

for all $0<\psi<1$, and all $0<\nu<\infty$. $\rho_0(\psi, \nu)$ is the R.H.S. of (4.3). Formula (4.15) can be utilized to obtain the solution of this functional equation. We present here the solution for the case of $\nu=1$ (prior negative exponential distribution of $\theta$).

When $\nu=1$ the $\gamma$-th fractile of geometric distribution $G(x|\psi, 1)$, say $k^0$, satisfies the inequality:

$$\frac{p}{c+p} \leqslant 1-\psi^{k^0-1} \text{ and } \frac{p}{c+p} > 1-\psi^{k^0}.$$

Hence, $\psi^{k^0+1} \leqslant c/c+p$ while $\psi^{k^0} > c/c+p$. We shall therefore approximate $\psi^{k^0+1}$ by $c/c+p$; i.e., $\psi^{k^0+1} \approx c/(c+p)$, and $k^0+1 \approx \log\,(c/c+p)/\log\,\psi$. By substituting this approximate value of $k^0+1$ in the R.H.S. of (4.3), we obtain:

$$\rho_0(\psi, 1) \approx c\,\frac{\log\left(\frac{c}{c+p}\right)}{\log\,\psi}, \text{ for } 0<\psi<1.$$

We therefore consider, for the case of $\nu=1$, the functional equation:

$$\rho_\infty(\psi, 1) = c\,\frac{\log\left(\frac{c}{c+p}\right)}{\log\,\psi} + (1-\psi)\sum_{x=0}^{\infty}\psi^x\rho_\infty\left(\frac{\psi}{1+\psi}, 1+x\right). \tag{5.2}$$

From (4.15), (5.2), and the relationship $\rho_\infty(\psi, 1) = \lim_{N\to\infty}\rho_{N-1}(\psi, 1)$, we obtain:

$$\rho_\infty(\psi, 1) = c\,\frac{\log\left(\frac{c}{c+p}\right)}{\log\,\psi} + \sum_{j=2}^{\infty}\psi_{j-1}\sum_{x=0}^{\infty}\rho_0(\psi_j, 1+x)(1-\psi_{j-1})^x, \tag{5.3}$$

where

$$\rho_0(\psi_j, 1+x) = (c+p)\prod_{i=0}^{x}(\tilde{k}_j+1+i)\psi_j^{\tilde{k}_j+1}(1-\psi_j)^{\nu-1}, \tag{5.4}$$

$$\tilde{k}_j = G_{\frac{p}{c+p}}(\psi_j, 1+x), \qquad x=0, 1, \ldots \tag{5.5}$$

and

$$\psi_j = \psi/(1+(j-1)\psi), \qquad j=1, 2, \ldots. \tag{5.6}$$

In a similar fashion, we can obtain an expansion for $\rho_\infty(\psi, \nu)$ in terms of the Bayes risk $\rho_0(\psi, \nu)$.

## 6. OPTIMAL SEQUENTIAL DETERMINATION OF STOCK LEVELS UNDER CONSTRAINT

The general solution of Section 4 was obtained under the assumption that the optimal stock level for any period is independent of the stock level at the previous period. This is in certain cases an unrealistic assumption, and we have to impose a constraint on the solution. The sufficient statistic for the decision concerning the optimal stock level at the beginning of the $n$-th period is the triplet $(\psi_n, \nu_n, Y_{n-1})$, where $Y_{n-1} \geq 0$ is the stock level at the *end* of the $(n-1)$st period. We thus have the constraint:

$$k_n^0(\psi_n, \nu_n, Y_{n-1}) \geq Y_{n-1}, \text{ for all } n = 1, 2, \ldots, N. \tag{6.1}$$

If we designate by $\rho_n(y|\psi_{N-n}, \nu_{N-n})$, $n = 0, 1, \ldots, N-1$, the Bayes risk function, given that at the end of the $(N-n-1)$st period the stock level is $Y_{N-n-1} = y$, then we have the following recursive relations for each $y \geq 0$,

$$\rho_0(y|\psi_N, \nu_N) = \inf_{k \geq y} \sum_{x=0}^{\infty} g(x|\psi_N, \nu_N) L(x; k) \tag{6.2}$$

and

$$\rho_n(y|\psi_{N-n}, \nu_{N-n}) = \inf_{k \geq y} \left\{ \sum_{x=0}^{\infty} g(x|\psi_{N-n}, \nu_n) \cdot \right. \tag{6.3}$$

$$\left. [L(x; k) + \rho_{n-1}(k-x)^+|\psi_{N-n+1}, \nu_n + x)] \right\}, \qquad n = 1, 2, \ldots, N-1.$$

The optimal *initial* stock level is $Y_0 = y^0$ which minimizes $\rho_{N-1}(y|\psi, \nu)$, i.e.,

$$\rho_{N-1}(y^0|\psi, \nu) = \inf_{0 \leq y < \infty} \rho_{N-1}(y|\psi, \nu). \tag{6.4}$$

This value depends, obviously, on the prior parameters $(\psi, \nu)$. The optimal stock level for the beginning of the 2nd period is the integer $k_2^0(\psi_2, \nu + X_1(y^0 - X_1)^+)$ which minimizes

$$\sum_{x=0}^{\infty} g(x|\psi_2, \nu + X_1)[L(x; k) + \rho_{N-3}((k-x)^+|\psi_3, \nu + X_1 + x)],$$

with respect to all intergers $k \geq (y^0 - X_1)^+$. In order to determine $k_2^0(\psi_2, \nu + X_1, (y^0 - X_1)^+)$ we need to know the function $\rho_{N-3}(y|\psi_3, \nu + x)$ for all $y = 0, 1, 2, \ldots$ and $x = 0, 1, 2, \ldots$

Thus, by the backward induction method of dynamic programming we first tabulate the functions $\rho_n(y|\psi_{N-n}, \nu_{N-n})$ for all $y = 0, 1, \ldots$, all $\nu_{N-n} = \nu, \nu + 1, \ldots$, and for $n = 0, 1, \ldots, N-1$. We then determine the optimal stock levels sequentially, as described above.

## 7. EMPIRIC BAYES APPROACH

According to the empiric Bayes approach, if the prior parameters $\psi$ and $\nu$ are unknown, we substitute for these prior parameters strongly consistent estimators, in order to obtain a procedure which approaches asymptotically (with probability 1) the Bayes sequential procedure. Accordingly, if the demand for parts is a random variable following the assumed Bayesian model, one can approximate the optimal Bayes design of stock levels after performing a large number of experiments, without committing the procedure to certain values of $\psi$ and $\nu$. We should remember that the true parameter values $\psi$ and $\nu$ are generally unknown, and the decision maker chooses their values according to "past experi-

ence" or according to some other "subjective" criterion. The empiric Bayes approach gives a method for incorporating the information accumulating in a way which yields good estimators of the unknown prior parameters $\psi$ and $\nu$. We show now how this objective can be obtained.

According to the Bayesian model, the observed random variables $X_1, X_2, \ldots$ (representing the demand values) are independent, and have an identical marginal negative-binomial distribution with parameters $\psi$ and $\nu$. Since all the moments of $G(x|\psi, \nu)$ exist, the sequences

$$\left\{\frac{1}{n}\sum_{i=1}^{n} X_i : n=1, 2, \ldots\right\}$$

and

$$\left\{\frac{1}{n}\sum_{i=1}^{n} X_i^2 : n=1, 2, \ldots\right\}$$

obey the Strong Law of Large Numbers and hence the mean of the first $n$ observations,

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n} X_i,$$

is a strongly consistent estimator of $E\{X\}=\nu\psi/(1-\psi)$. Similarly, the sample variance

$$\hat{\sigma}_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$$

is a strongly consistent estimator of var $\{X\}=\nu\psi/(1-\psi)^2$. Thus, we estimate $\psi$ by

$$\hat{\psi}=\left(1-\frac{\bar{X}_n}{\hat{\sigma}_n^2}\right)^+.$$

We notice that for small values of $n$, $\hat{\sigma}_n^2$ might be, with positive probability, smaller than $\bar{X}_n$. This is the reason for defining (7.1) to be the positive part of $1-\bar{X}_n/\hat{\sigma}_n^2$; however, as $n\to\infty$ we have $\bar{X}_n\to E\{X\}$ a.s. and $\hat{\sigma}_n^2\to$ Var $\{X\}$ a.s. Hence, $\hat{\psi}\to\psi$ a.s. By similar arguments we construct the estimator of $\nu$, namely:

$$\hat{\nu}=\bar{X}_n(1-\hat{\psi})/\hat{\psi}, \text{ if } \hat{\psi}>0. \tag{7.2}$$

This estimator converges a.s. to $\nu$. For small values of $n$, $\hat{\psi}$ could be 0 with positive probability. In these cases $\hat{\nu}$ is not defined.

Whenever $\hat{\psi}>0$ one can use the estimators (7.1) and (7.2) to obtain a consistent estimator of the $\gamma$-th fractile of $G(X|\psi_n, \nu_n)$. This estimator is defined as:

$$\hat{k}_n^0(\hat{\psi}_n, \hat{\nu}_n)=\text{least integer } k\geq 0 \text{ such that } I_{1-\hat{\psi}_n}(\hat{\nu}_n, k+1)\geq\gamma,$$

where $\hat{\psi}_n=\hat{\psi}/(1+(n-1)\hat{\psi})$ and $\hat{\nu}_n=\hat{\nu}+S_{n-1}$.

We notice that, as $n\to\infty$, we have asymptotically, $k_n^0(\hat{\psi}_n, \hat{\nu}_n)\approx k_n^0(\psi_n, \nu_n)$ a.s. As seen in Table 1, Section 3, the values of $k_n^0(\psi_n, \nu_n)$ are quite sensitive to variations in $\nu_n$, and for large values of $\nu_n$ they are sensitive also to variations in $\psi_n$. Thus, the Bayes sequential rule might be worthless if we do not have good estimates of $\nu_n$ and of $\psi_n$. It is therefore suggested not to use the Bayes sequential procedure if good estimates of $\psi_n$ and $\nu_n$ are not available. In these cases, one can use, for example, a tolerance limit approach to control the stock level. These tolerance limits can be adjusted sequentially as shown in [7].

## 8. APPENDIX

### 8.1 Proof of the relationship

$$(8.1)\qquad (1-\psi)^{\nu}\sum_{x=0}^{k}\frac{\Gamma(\nu+x)}{\Gamma(\nu)\Gamma(x+1)}\psi^{x}=I_{1-\psi}(\nu,\ k+1),$$

for all $0<\psi<1$, all $0<\nu<\infty$ and all $k=0,\ 1,\ \ldots$.

**Proof**

Write the geometric distribution $G(k|\psi^*,\ \nu)$ on the L.H.S. of (8.1) as a mixture of the Poisson distribution

$$P(k;\ \theta)=e^{-\theta}\sum_{x=0}^{k}\frac{\theta^{x}}{x!}$$

with respect to the gamma distribution $\mathcal{G}(1/\psi,\ \nu)$ for $\theta$, where $\psi^*=\psi/(1+\psi)$. That is,

$$(8.2)\qquad G(k|\psi,\ \nu)=\frac{\psi^{-\nu}}{\Gamma(\nu)}\int_{0}^{\infty}e^{-\theta/\psi}\theta^{\nu-1}P(k;\ \theta)d\theta.$$

We express now the Poisson $P(k;\ \theta)$ in terms of the incomplete gamma function ratio, i.e.,

$$(8.3)\qquad P(k;\ \theta)=\frac{\theta^{k+1}}{\Gamma(k+1)}\int_{1}^{\infty}t^{k}e^{-\theta t}dt.$$

Substituting (8.3) in (8.2) and interchanging the order of integration (Fubini's theorem holds) we obtain:

$$(8.4)\qquad G(k|\psi,\ \nu)=\frac{\psi^{-\nu}}{\Gamma(\nu)\Gamma(k+1)}\int_{1}^{\infty}dt\cdot t^{k}\left\{\int_{0}^{\infty}\theta^{\nu+k}e^{-\left(\frac{1}{\psi}+t\right)\theta}\,d\theta\right\}$$

$$=\frac{\psi^{-\nu}}{B(\nu,\ k+1)}\int_{1}^{\infty}t^{k}\left(\frac{1}{\psi}+t\right)^{-(\nu+k+1)}dt$$

$$=\frac{\psi^{k}}{B(\nu,\ k+1)}\int_{1}^{\infty}t^{k}(1+\psi t)^{-(\nu+k+1)}dt.$$

Making the transformation $\mu=(1+\psi t)^{-1}$, we easily prove that

$$(8.5)\qquad G(k|\psi,\ \nu)=\frac{1}{B(\nu,\ k+1)}\int_{0}^{(1+\psi)^{-1}}\mu^{\nu-1}(1-\mu)^{k}\,d\mu$$

$$\equiv I_{(1+\psi)^{-1}}(\nu,\ k+1).$$

Finally, since $(1+\psi)^{-1}=1-\psi^*$, we obtain (8.1) by replacing $\psi^*$ by $\psi$.

### 8.2 Proof of approximation (4.3)

The risk function under a negative-binomial distribution $G(X|\psi,\ \nu)$ is:

$$(8.6)\qquad R(k;\ \psi,\ \nu)=\sum_{x=0}^{\infty}L(k;\ x)g(x|\psi,\ \nu)$$

$$=(c+p)kG(k|\psi,\ \nu)-pk$$

$$-(c+p)\sum_{x=0}^{k}xg(x|\psi,\ \nu)+p\sum_{x=0}^{\infty}xg(x|\psi,\ \nu).$$

We are interested at $\rho_0(\psi, \nu) \equiv R(k^0, \psi, \nu)$, where $k^0$ is the $(p/c+p)$th fractile of $G(X|\psi, \nu)$. Hence, $G(k^0|\psi, \nu) \cong p/c+p$ (actually $G(k^0|\psi, \nu) \geq p/c+p$). Hence,

$$\rho_0(\psi, \nu) \approx -(c+p) \sum_{x=0}^{k^0} xg(x|\psi, \nu) + p \sum_{x=0}^{\infty} xg(x|\psi, \nu). \tag{8.7}$$

We notice that

$$\sum_{x=0}^{\infty} xg(x|\psi, \nu)$$

is the expected value of a negative binomial r.v. Hence, as it is simple to verify

$$\sum_{x=0}^{\infty} xg(x|\psi, \nu) = \nu \frac{\psi}{1-\psi}, \qquad \text{for all } 0 < \psi < 1. \tag{8.8}$$

Let

$$M(k|\psi, \nu) = \sum_{x=0}^{k} xg(x|\psi, \nu), \qquad k = 0, 1, \ldots. \tag{8.9}$$

Obviously, $M(\infty|\psi, \nu) = \nu \cdot \psi/(1-\psi)$ and $M(0|\psi, \nu) = 0$. Now for all $k \geq 1$ we have:

$$\begin{aligned} M(k|\psi, \nu) &= \sum_{x=1}^{k} \frac{\Gamma(\nu+x)}{\Gamma(\nu)\Gamma(x)} \psi^x (1-\psi)^\nu \\ &= \psi \sum_{y=0}^{k-1} y \frac{\Gamma(y+\nu)}{\Gamma(\nu)\Gamma(y+1)} \psi^y (1-\psi)^\nu + \psi\nu \sum_{y=0}^{k-1} \frac{\Gamma(y+\nu)}{\Gamma(\nu)\Gamma(y+1)} \psi^y (1-\psi)^\nu, \end{aligned} \tag{8.10}$$

or

$$M(k|\psi, \nu) = \psi \left[ M(k|\psi, \nu) - \frac{\Gamma(k+\nu)}{\Gamma(\nu)\Gamma(k)} \psi^k (1-\psi)^\nu \right] + \psi\nu G(k-1|\psi, \nu). \tag{8.11}$$

Hence,

$$\frac{1-\psi}{\psi} M(k|\psi, \nu) = -\frac{\Gamma(k+\nu)}{\Gamma(k)\Gamma(\nu)} \psi^k (1-\psi)^\nu \left(1 + \frac{\nu}{k}\right) + G(k|\psi, \nu). \tag{8.12}$$

Hence, for all $k \geq 1$

$$M(k|\psi, \nu) = \frac{\psi}{1-\psi} \left\{ \nu G(k|\psi, \nu) - \frac{\Gamma(k+\nu+1)}{\Gamma(\nu)\Gamma(k+1)} \psi^k (1-\psi)^\nu \right\}. \tag{8.13}$$

Finally, substituting (8.8) and (8.13) in (8.7) with $k = k^0$, we obtain:

$$\begin{aligned} \rho_0(\psi, \nu) &\approx -(c+p) M(k^0|\psi, \nu) + p\nu \frac{\psi}{1-\psi} \\ &\cong (c+p) \frac{\Gamma(k^0+\nu+1)}{\Gamma(\nu)\Gamma(k^0+1)} \psi^{k^0+1} (1-\psi)^{\nu-1}. \end{aligned} \tag{8.14}$$

This proves (4.3).

## REFERENCES

[1] Harter, L. H., *New Tables of the Incomplete Gamma-Function Ratio and of Percentage Points of the Chi-Square and Beta Distributions*, Aerospace Research Laboratories, U.S. Air Force (1964).

[2] Karlin, S., "Dynamic Inventory Policy with Varying Stochastic Demands," *Management Science*, **6**, 231–258 (1960).

[3] Pearson, K., *Tables of the Incomplete Beta-Function* (Cambridge University Press for the Biometrika Trustees, Cambridge, 1934).

[4] Raiffa, H., and R. Schlaifer, "*Applied Statistical Decision Theory*," Graduate School of Business Administration, Harvard University, Boston (1961).

[5] Scarf, H., "Bayes Solutions of the Statistical Inventory Problem," *Annals of Mathematical Statistics*, **30**, 490–508 (1959).

[6] Veinott, A. F., Jr., "The Status of Mathematical Inventory Theory," *Management Science*, **12**, 745–777 (1966).

[7] Zacks, S., "Uniformly Most Accurate Upper Tolerance Limits in the Poisson Case, and Its Application to Inventory Control," T.R. #2, NSF Project GP–9007, Dept. of Mathematics and Statistics. University of New Mexico (1968).

* * *

# A BRANCH AND BOUND ALGORITHM FOR ALLOCATION PROBLEMS IN WHICH CONSTRAINT COEFFICIENTS DEPEND UPON DECISION VARIABLES

Donald Gross

*School of Engineering and Applied Science*
*The George Washington University*
*Washington, D.C.*

and

Richard M. Soland

*Advanced Research Department*
*Research Analysis Corporation*
*McLean, Virginia*

**ABSTRACT**

A branch and bound algorithm is developed for a class of allocation problems in which some constraint coefficients depend on the values of certain of the decision variables. Were it not for these dependencies, the problems could be solved by linear programming. The algorithm is developed in terms of a strategic deployment problem in which it is desired to find a least-cost transportation fleet, subject to constraints on men/materiel requirements in the event of certain hypothesized contingencies. Among the transportation vehicles available for selection are aircraft which exhibit the characteristic that the amount of goods deliverable by an aircraft on a particular route in a given time period (called aircraft productivity and measured in kilotons/aircraft/month) depends on the ratio of type 1 to type 2 aircraft used on that particular route. A model is formulated in which these relationships are first approximated by piecewise linear functions. A branch and bound algorithm for solving the resultant nonlinear problem is then presented: the algorithm solves a sequence of linear programming problems. The algorithm is illustrated by a sample problem and comments concerning its practicality are made.

## 1. INTRODUCTION

Many allocation problems have been formulated and solved as linear programming problems. A general formulation is: find $x$ to minimize $cx$ subject to $Ax \geq b$ and $x \geq 0$. Here $x$ is an allocation or decision vector, $c$ is a constant cost vector, and $b$ is a constant vector of requirements and/or availabilities. $A$ is a constant matrix whose elements relate the allocation variables to the requirements. Each element $a_{ij}$ of $A$ may be generally thought of as a productivity or coefficient of effectiveness. In this paper, we consider a class of problems in which it is assumed that some of these productivities are not constant, but instead depend upon the allocation variables. A branch and bound algorithm is developed to solve such problems. The exact nature of these problems and the algorithm for solving them are developed in terms of a strategic deployment problem because (1) a strategic deployment formulation facilitates and clarifies the development of the methodology, (2) the strategic deployment model is in itself of interest, and (3) the model is potentially useful in the analysis of strategic deployment problems.

Reference [3] describes a least-cost, strategic deployment linear programming model, which, when faced with certain possible contingencies, indicates the manner in which materiel is to be deployed to meet requirements imposed by these contingencies. This model assumes that aircraft productivity, that is, the tonnage per month a given type of aircraft can deliver from a particular source to a particular destination, is a constant. It has been argued by personnel experienced with aircraft loading that this assumption is not always realistic since aircraft productivity on a given route may strongly depend on the aircraft mix, that is, on the ratio of the numbers of one type of aircraft to another used on the route. This dependency has been observed in practice (see Table 1 and Figure 1) and might be explained by the following logic: The greater the number of, say, type 2 aircraft relative to type 1 aircraft used, the more efficiently can the type 1 aircraft be loaded (because more items not efficiently carried by type 1 aircraft are shunted to type 2 aircraft), and hence the greater the productivity of the type 1 aircraft. This paper describes an algorithm for a model incorporating a piecewise linear relationship between productivity and aircraft mix. In Section 2 we describe the model with constant productivity in terms of an example from [8]. In Section 3 we examine aircraft productivities as functions of aircraft mix, approximate these functions by two linear pieces, and expand the model to include this generalization. This leads to a mathematical programming problem that is only piecewise linear, and in Section 4 we develop a branch and bound algorithm to solve this problem. A numerical example is given in Section 5, and in Section 6 we comment on the practicality of this approach.

TABLE 1. *Productivity of mixes of C–141 and C–5A aircraft on a route from Forbes Air Force Base, Kans., to Khorat, Thailand*[a]

(Data obtained from the Lockheed-Georgia Company, July 1967)[b]

| Mix | | Actual sorties | | Average payload, lb | | Average cycle time, hr | | Productivity of 1000 tons/aircraft/month | |
|---|---|---|---|---|---|---|---|---|---|
| C–141/C–5A | C–5A/C–141 | C–5A | C–141 | C–5A | C–141 | C–5A | C–141 | C–5A | C–141 |
| 0/1 (0) | 1/0 (∞) | 886 | ..... | 235,943 | ..... | 96.845 | ..... | 0.887 | ..... |
| 1/2 (0.5) | 2/1 (2) | 781 | 390 | 234,784 | 66,244 | 97.036 | 95.405 | 0.871 | 0.250 |
| 1/1 (1) | 1/1 (1) | 700 | 701 | 234,275 | 64,463 | 96.873 | 95.391 | 0.870 | 0.243 |
| 2/1 (2) | 1/2 (0.5) | 580 | 1159 | 236,784 | 61,980 | 97.037 | 95.384 | 0.873 | 0.234 |
| 5/1 (5) | 1/5 (0.2) | 384 | 1949 | 244,594 | 59,058 | 97.493 | 95.233 | 0.903 | 0.224 |
| [c] 1/0 (∞) | [c] 0/1 (0) | ..... | ..... | ..... | [c] 54,000 | ..... | [c] 95.3 | ..... | [c] 0.204 |

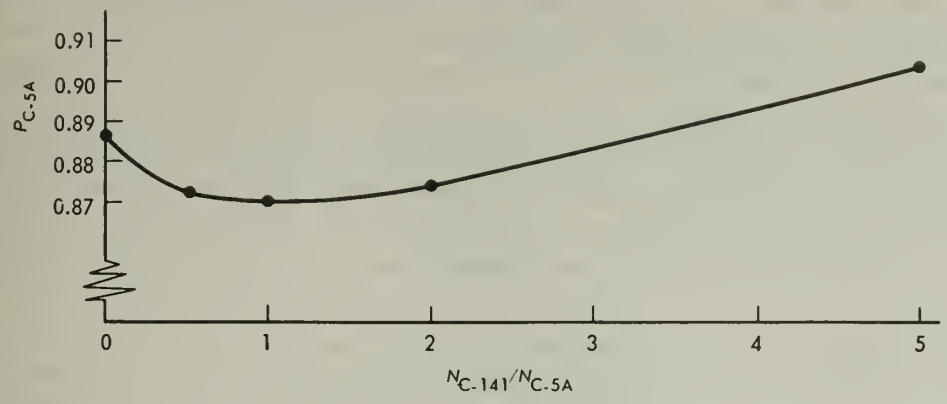
[a] Total weight of force A, 209,198,892 lb; weight of vehicles outsize to C–141, 52,314,277 lb; mission, Forbes Air Force Base, Kans., to Khorat, Thailand; distance, approximately 8540 NM; assumed aircraft utilization: 10 hr/day.

[b] The authors wish to express their appreciation to Mr. J. D. Grow and Mr. T. G. Greenlee of the Lockheed-Georgia Company for providing these data.

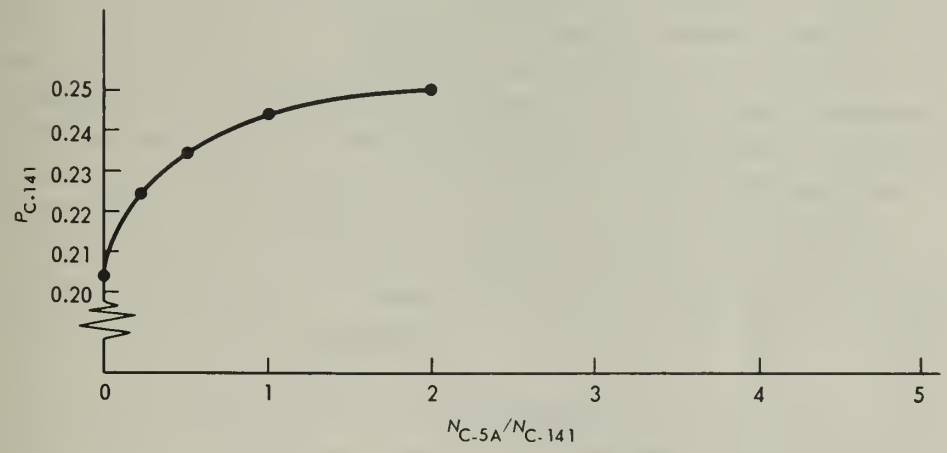
[c] Data obtained from earlier Lockheed-Georgia computer runs of June 1966.

## 2. CONSTANT PRODUCTIVITY LINEAR PROGRAMMING MODEL

The following example, a modification of that found in [8], illustrates the major characteristics of the type of strategic deployment model with which we are concerned. Figure 2 describes the distribution network. Country $A$ is responsible for deploying materiel to countries $B$, $C$, $D$, and $E$ in case these are involved in certain possible contingencies. There are three contingencies considered: (1) $B$ and $C$ must be supplied simultaneously, (2) $B$ and $D$ must be supplied simultaneously, and (3) $E$ must be supplied. Note that if country $C$ is involved in any action, country $B$ will also be involved, i.e., $B$ and $C$ become involved simultaneously. The same holds true for countries $D$ and $B$. If a contingency involving




a.




b.

FIGURE 1. Productivity of C–5A and C–141 as a function of mix.

$E$ occurs, only country $E$ becomes involved. If any country ($B$, $C$, $D$, or $E$) is involved in an action, there are specific materiel requirements which must be met. Country $A$ has a choice of how to meet these requirements. In Figure 2, a/s indicates the possibilities of transporting goods by air or sea (this

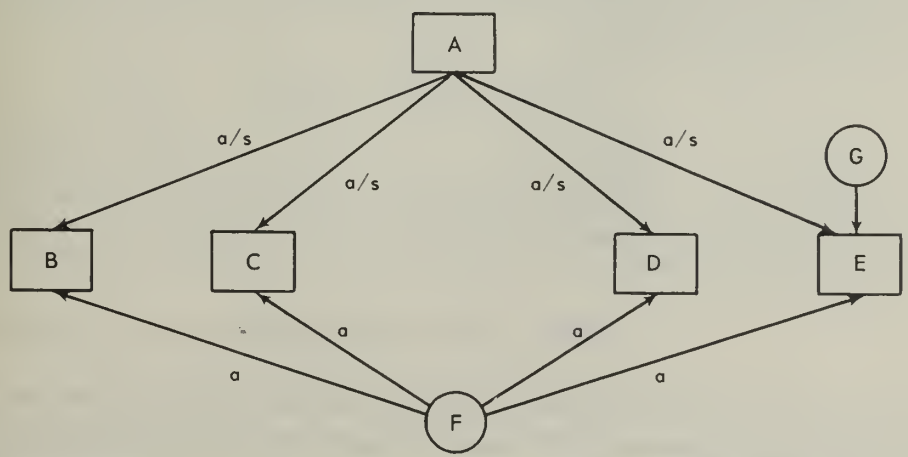



FIGURE 2. Deployment network for Example I a/s, air or sea; a, air.

example assumes two types of aircraft and one type of ship available). In addition to direct air or sea transportation to the countries, an option of prepositioning materiel at sites $F$ and/or $G$ exists. Materiel at site $F$ can then be transported to the various countries, but by aircraft only. All materiel prepositioned at site $G$ is for the exclusive use of country $E$, and site $G$ is physically located at country $E$ so that no transshipment is required. For this example only one time period is considered and all requirements must be met by the end of the period. (The methodology developed in this paper would also apply to multiperiod models.) Subject to the constraints implied by the above description, country $A$ wishes to procure numbers of aircraft and ships and establish bases at $F$ and/or $G$ so as to minimize the cost of these items.

Figure 3 shows the structure of the linear programming strategic deployment model outlined above; aircraft productivities are assumed constant. There are two types of variables considered in the model. Those beginning with an $S$ designation are called system variables, while those beginning with an $N$ designation are referred to as deployment variables. The variable PEV will be explained later. System variables $S_1$ through $S_5$, respectively, represent the total numbers of type 1 aircraft, type 2 aircraft, and ships (only one type of ship is assumed in this example) in the system, the total tonnage prepositioned at site $F$, and the total tonnage prepositioned at site $G$. The deployment variables represent the number of vehicles used on particular routes and have three subscripts. The first of these denotes the vehicle type (1 represents type 1 aircraft, 2 represents type 2 aircraft, and 3 represents ship). The second and third subscripts represent source and destination, respectively. For example $N_{1AB}$ is the number of type 1 aircraft used to deploy materiel from $A$ to $B$; the number of ships used to deploy materiel from $A$ to $D$ is $N_{3AD}$; $N_{1FD}$ is the number of type 1 aircraft used to deploy materiel prepositioned at site $F$ to $D$; and so on. There are no deployment variables needed for prepositioned materiel at site $G$.

The first row, marked COST, is the objective function. Note that only the system variables appear in the objective function; they entail budgetary or programming costs. Since contingencies may never occur, costs of deploying materiel may never be realized and are often not considered, as is the case in this example. For a further discussion of deployment or utilization costs as they are often called, see [7].

The first six constraints, rows $S1L$ to $S4S5L$ inclusive, are limits on availabilities of the system variables. A "+" in the leftmost column indicates that the constraint is an inequality of the less than or equal to ($\leq$) type. A "−" indicates an inequality of the greater than or equal to ($\geq$) type. Constraints $S1BC$ to $S4E$ inclusive relate the deployment variables to the system variables for each of the contingencies. For example, $S1BC$ gives

$$S_1 \geq N_{1AB} + N_{1FB} + N_{1AC} + N_{1FC} \tag{1}$$

which requires the number of type 1 aircraft in the system to be greater than or equal to the number utilized for the $BC$ contingency. Constraints $S2BC$ and $S3BC$ are similar for type 2 aircraft and ships, respectively. Constraint $S4BC$ requires that total prepositioned tonnage at $F$ be greater than or equal to that used for contingency $BC$, that is

$$S_4 \geq XN_{1FB} + AN_{2FB} + XN_{1FC} + XN_{2FC}, \tag{2}$$

where the coefficients are aircraft productivities for the respective routes. (All aircraft productivities are circled in Figure 3.) Constraints $S1BD$ through $S4E$ inclusive are similar to those described above, but apply to the other two contingencies, that is, contingency $BD$ and contingency $E$.

| Constraint name | Variable | | | | | | | | | | | | | | | | | | | | | | | | | RHS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | | |
| | | | | | | 1 | 2 | 3 | 1 | 2 | 1 | 2 | 3 | 1 | 2 | 1 | 2 | 3 | 1 | 2 | 1 | 2 | 3 | 1 | 2 | P | R |
| | S | S | S | S | S | A | A | A | F | F | A | A | A | F | F | A | A | A | F | F | A | A | A | F | F | E | H |
| | 1 | 2 | 3 | 4 | 5 | B | B | B | B | B | C | C | C | C | C | D | D | D | D | D | E | E | E | E | E | V | S |
| Cost | X | X | X | X | A | | | | | | | | | | | | | | | | | | | | | –X | |
| + S1L | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | Y |
| + S2L | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | Y |
| + S3L | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | Z |
| + S4L | | | | 1 | | | | | | | | | | | | | | | | | | | | | | | Y |
| + S5L | | | | | 1 | | | | | | | | | | | | | | | | | | | | | | Y |
| + S4S5L | | | | 1 | 1 | | | | | | | | | | | | | | | | | | | | | | Y |
| + S1BC | –1 | | | | | 1 | | | 1 | | 1 | | | 1 | | | | | | | | | | | | | |
| + S2BC | | –1 | | | | | 1 | | | 1 | | 1 | | | 1 | | | | | | | | | | | | |
| + S3BC | | | –1 | | | | | 1 | | | | | 1 | | | | | | | | | | | | | | |
| + S4BC | | | | –1 | | | | | (X) | (A) | | | | (X) | (X) | | | | | | | | | | | | |
| + S1BD | –1 | | | | | 1 | | | 1 | | | | | | | 1 | | | 1 | | | | | | | | |
| + S2BD | | –1 | | | | | 1 | | | 1 | | | | | | | 1 | | | 1 | | | | | | | |
| + S3BD | | | –1 | | | | | 1 | | | | | | | | | | 1 | | | | | | | | | |
| + S4BD | | | | –1 | | | | | (X) | (A) | | | | | | | | | (X) | (A) | | | | | | | |
| + S1E | –1 | | | | | | | | | | | | | | | | | | | | 1 | | | 1 | | | |
| + S2E | | –1 | | | | | | | | | | | | | | | | | | | | 1 | | | 1 | | |
| + S3E | | | –1 | | | | | | | | | | | | | | | | | | | | 1 | | | | |
| + S4E | | | | –1 | | | | | | | | | | | | | | | | | | | | (A) | (B) | | |
| – RBL | | | | | | (X) | (A) | X | (X) | (A) | | | | | | | | | | | | | | | | | Y |
| + PBL | | | | | | | | X | | | | | | | | | | | | | | | | | | | Y |
| – RCL | | | | | | | | | | | (A) | (A) | X | (X) | (X) | | | | | | | | | | | | Y |
| + PCL | | | | | | | | | | | | | X | | | | | | | | | | | | | | Y |
| – RDL | | | | | | | | | | | | | | | | (A) | (B) | X | (X) | (A) | | | | | | | Y |
| + PDL | | | | | | | | | | | | | | | | | | X | | | | | | | | | Y |
| – REL | | | | | 1 | | | | | | | | | | | | | | | | (A) | (B) | X | (A) | (B) | | Y |
| + PEL | | | | | | | | | | | | | | | | | | | | | | | X | | | | Y |
| + APT | –1 | –A | | | | | | | | | | | | | | | | | | | | | | | | 1 | |
| + PTRQ | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | Y |

| Legend | | | | | | |
|---|---|---|---|---|---|---|
| + ≤ inequality | ≤ | Symbol | > | ≤ | Symbol | > |
| 0 = equality | 9999 | W | 1000 | 1.0 | A | 0.1 |
| – ≥ inequality | 1000 | Z | 100 | 0.1 | B | 0.01 |
| | 100 | Y | 10 | 0.01 | C | 0.001 |
| | 10 | X | 1 | 0.001 | D | 0.0001 |
| | | | | 0.0001 | E | 0.00001 |
| | | | | 0.00001 | F | 0.000001 |

FIGURE 3. Strategic deployment Example 1, constant aircraft productivity (circled symbols are aircraft productivities).

Constraint *RBL* governs the requirements of country *B* as follows:

$$XN_{1AB} + AN_{2AB} + XN_{3AB} + XN_{1FB} + AN_{2FB} \geq Y, \tag{3}$$

where, again, the coefficients are productivities of the vehicles, which when multiplied by the number of vehicles yield tonnage deployed. Constraints *RCL*, *RDL*, and *REL* are similar for countries *C*,

*D*, and *E*, respectively. Constraints *PBL*, *PCL*, *PDL*, and *PEL* restrict total tonnage deployed by ship due to port limitations.

The last two constraints are connected with the variable *PEV*, which represents the peacetime value of an aircraft. Some aircraft can be useful for peacetime resupply in that when considering both transportation and inventory costs, it is more economical to transport certain material by air than by sea (see [7]). Since not all of the aircraft required for strategic deployment may be usable during peacetime, *PEV* represents the aircraft that can be utilized. Constraint *APT* relates the relative peacetime contribution of the two types of aircraft and requires that those aircraft utilized for peacetime not exceed the total aircraft in the system. Constraint *PTRQ* limits the number of aircraft usable for peacetime resupply. Note that there is a cost credit for *PEV* in the objective function.

## 3. EXPANSION OF THE MODEL TO INCLUDE PRODUCTIVITY AS A FUNCTION OF MIX

Analysis of productivity data, see Table 1 and Figure 1, has suggested that the productivity of, say, type 1 aircraft on a particular route depends on the numbers of type 1 and type 2 aircraft on that route through the ratio of type 2 to type 1 aircraft. Figure 4 shows the general form of this dependence. The relationship assumed here has a diminishing rate of productivity increase with increasing mix ratio. This is in general agreement with the data shown in Figure 1 except for a small portion of the curve shown in Figure 1a. This discrepancy may be due to the idiosyncrasies of the particular loading program used.

One approach to solving the problem with nonconstant productivities is to approximate the productivity data by functions that exhibit the characteristics shown in Figure 4 (e.g., exponential or hyperbolic functions) and use a nonlinear programming technique. Unfortunately, for the type of problems under study (e.g., that discussed in Section 2) this usually leads to nonconvex programming problems, for which global solutions cannot be guaranteed. Computational experience using exponential functions for the productivities and the sequential unconstrained minimization technique [2] proved unsatisfactory.

The approach taken here is to approximate the dependence of productivity on aircraft mix by piecewise linear functions. It will be shown that this results in the necessity to effectively solve a number of linear programming problems; a branch and bound algorithm is developed to do this efficiently. In order to describe the model we introduce the following notation:

$N_{ijk}$ = Number of aircraft type $i$ used to supply country $k$ from source $j$ (route $jk$), $i = 1, 2$.

$P_{ijk}$ = Productivity of aircraft type $i$ on route $jk$.

$a_{ijk}$ = Minimum productivity of aircraft type $i$ on route $jk$ (aircraft type $i$ used alone).

$b_{ijk}$ = Maximum productivity of aircraft type $i$ on route $jk$.

Consider a particular route $jk$. We approximate by the piecewise linear function shown in Figure 5 the relationship between productivity and aircraft mix given in Figure 4. We will assume that $L_{1jk}L_{2jk} > 1$ (see Figure 5). This implies three different possibilities for the two productivity functions:

(i) If $L_{2jk}^{-1} \leq N_{2jk}/N_{1jk} \leq L_{1jk}$, line segments ① and ③ of Figure 4 are valid for $P_{1jk}$ and $P_{2jk}$, respectively.

(ii) If $N_{2jk}/N_{1jk} \leq L_{2jk}^{-1}$, line segments ① and ④ of Figure 4 are valid for $P_{1jk}$ and $P_{2jk}$, respectively.

(iii) If $N_{2jk}/N_{1jk} \geq L_{1jk}$, line segments ② and ③ of Figure 4 are valid for $P_{1jk}$ and $P_{2jk}$, respectively.

(Note that if $L_{1jk}L_{2jk} < 1$ we will get a somewhat different set of three possibilities. If $L_{1jk}L_{2jk} = 1$, we get only two possibilities; however, the methodology that follows will still apply, requiring only slight

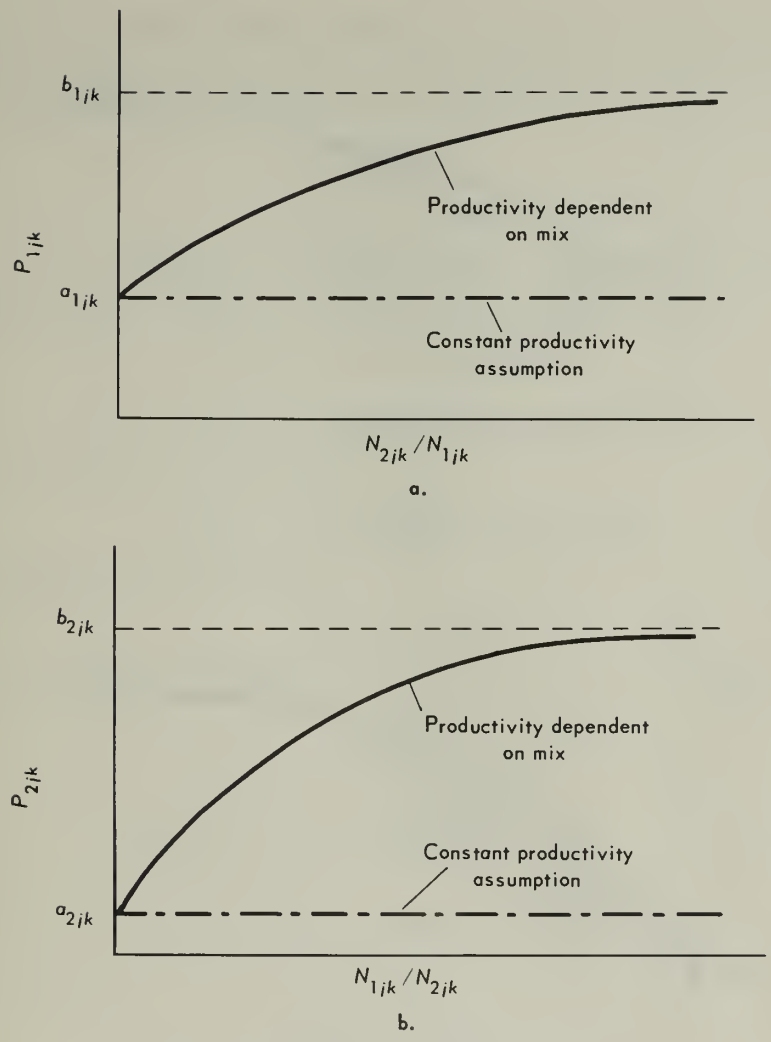



FIGURE 4. Productivity as a function of aircraft mix.

adjustments in certain places.) The specific $L$ values used should be chosen on the basis of the observed productivity data. Indeed, the $a$, $b$, and $L$ values of Figure 5 should be chosen together on this basis. One procedure would be to find the $a$, $b$, and $L$ values that minimize the sum of squared deviations or sum of absolute deviations between the piecewise linear function and the data points. One consideration is that choosing the values of $a$, $b$, and $L$ so that the piecewise linear function lies below the data points would provide a slightly pessimistic estimate of the productivity function. Such pessimistic estimates would lead to a slightly conservative solution to the overall problem.

To consider the effect of such piecewise linear functions recall that productivities entered into the requirement constraints which were of the form given by Eq. (3). The general form is

$$\sum_j [P_{1jk}N_{1jk} + P_{2jk}N_{2jk}] + T_k \geq R_k, \tag{4}$$

where $R_k$ is the total tonnage required at country $k$, and $T_k$ represents tonnage deployed to country $k$ by vehicles other than aircraft. (In the previous example, the only other vehicles we have are ships.) For the constant productivity case, $P_{1jk} = a_{1jk}$ and $P_{2jk} = a_{2jk}$ yielding constraints of the form

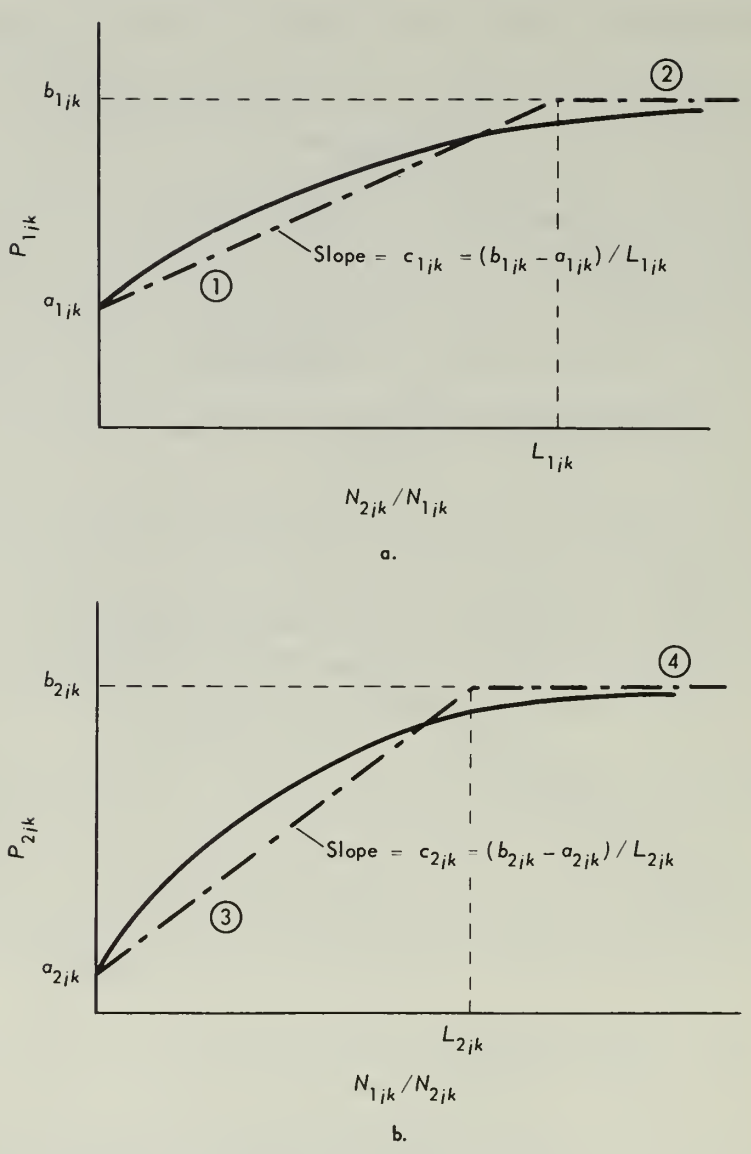



FIGURE 5. Piecewise linear approximation to productivity curves.

$$\sum_j [a_{1jk}N_{1jk} + a_{2jk}N_{2jk}] + T_k \geq R_k. \tag{5}$$

(See Figure 3 constraints *RBL* to *REL* inclusive.)

For the piecewise linear productivity functions the form of the requirement constraint given by Eq. (4) will depend on which of the above three cases applies. For case (i), using ① and ③ of Figure 5, we have

$$P_{1jk} = a_{1jk} + c_{1jk}(N_{2jk}/N_{1jk}), \tag{6}$$

$$P_{2jk} = a_{2jk} + c_{2jk}(N_{1jk}/N_{2jk}). \tag{7}$$

By substituting Eqs. (6) and (7) into (4), the requirement constraint becomes

$$(a_{1jk} + c_{2jk})N_{1jk} + (a_{2jk} + c_{1jk})N_{2jk} + T'_k + T_k \geq R_k, \tag{8}$$

where $T'_k$ represents tonnage deployed by aircraft via other routes to country $k$ and $T_k$, as before, represents tonnage deployed to $k$ via other vehicles. In order to insure that the mix ratios are in the region indicated in case (i) the following constraints must be added:

$$N_{2jk} \leq L_{1jk} N_{1jk}, \tag{9}$$

$$N_{1jk} \leq L_{2jk} N_{2jk}. \tag{10}$$

(Note that the same type of analysis would apply to the tonnage deployed via other aircraft routes, represented by the $T'_k$ term.) For case (ii), using ① and ④ of Figure 5,

$$P_{1jk} = a_{1jk} + c_{1jk}(N_{2jk}/N_{1jk}), \tag{11}$$

$$P_{2jk} = b_{2jk}. \tag{12}$$

Substituting Eqs. (11) and (12) into (4), we get

$$a_{1jk} N_{1jk} + (b_{2jk} + c_{1jk}) N_{2jk} + T'_k + T_k \geq R_k. \tag{13}$$

The constraint required to ensure case (ii) is

$$N_{1jk} \geq L_{2jk} N_{2jk}. \tag{14}$$

For case (iii), using ② and ③ of Figure 5,

$$P_{1jk} = b_{1jk}, \tag{15}$$

$$P_{2jk} = a_{2jk} + c_{2jk}(N_{1jk}/N_{2jk}), \tag{16}$$

and Eq. (4) becomes

$$(b_{1jk} + c_{2jk}) N_{1jk} + a_{2jk} N_{2jk} + T'_k + T_k \geq R_k; \tag{17}$$

the added constraint is

$$N_{2jk} \geq L_{1jk} N_{1jk}. \tag{18}$$

In each of the three cases, all constraints are linear so that for a particular case we still have a linear programming problem. Thus, if there are a total of $r$ possible aircraft routes for the entire problem, since we have three cases for each route and the $r$ routes can be considered independently, there are a total of $3^r$ different cases for the entire problem. Therefore, what we have are $3^r$ linear programming problems to solve, each having the same objective function, but with different coefficients in certain of the requirement constraints plus added constraints to insure the proper region for the mix ratios as dictated by the particular case on each route. Letting $Z_j^*$ be the optimum value of the objective function for the $j^{\text{th}}$ problem ($j = 1,2, \ldots, 3^r$), suppose $Z_s^* = \min_j Z_j^*$. Then the optimal solution to the piecewise linear model is the solution to problem $s$.

Figure 6 shows the linear programming format for the $3^r$ problems. The values represented by the check marks depend on which of the $3^r$ problems is being solved. For any particular case, some of the check marks in the added constraint block may be zero, since the entire constraint set may not apply. In some situations it may be practicable to solve all $3^r$ problems. Usually it will not be practicable, and in the next section we present a branch and bound procedure for finding the optimal solution while solving far fewer problems.

| Constraint Name | Variable | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | N 1 | N 2 | N 3 | N 1 | N 2 | N 1 | N 2 | N 3 | N 1 | N 2 | N 1 | N 2 | N 3 | N 1 | N 2 | N 1 | N 2 | N 3 | N 1 | N 2 | P | R H S |
| | S 1 | S 2 | S 3 | S 4 | S 5 | A B | A B | A B | F B | F B | A C | A C | A C | F C | F C | A D | A D | A D | F D | F D | A E | A E | A E | F E | F E | E V | |
| Cast | X | X | X | X | A | | | | | | | | | | | | | | | | | | | | | –X | |
| + S1L | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | Y |
| + S2L | | 1 | | | | | | | | | | | | | | | | | | | | | | | | | Y |
| + S3L | | | 1 | | | | | | | | | | | | | | | | | | | | | | | | Z |
| + S4L | | | | 1 | | | | | | | | | | | | | | | | | | | | | | | Y |
| + S5L | | | | | 1 | | | | | | | | | | | | | | | | | | | | | | Y |
| + S4S5L | | | | 1 | 1 | | | | | | | | | | | | | | | | | | | | | | Y |
| + S1BC | –1 | | | | | 1 | | | 1 | | 1 | | | 1 | | | | | | | | | | | | | |
| + S2BC | | –1 | | | | | 1 | | | 1 | | 1 | | | 1 | | | | | | | | | | | | |
| + S3BC | | | –1 | | | | | 1 | | | | | 1 | | | | | | | | | | | | | | |
| + S4BC | | | | –1 | | | | | √ | √ | | | | √ | √ | | | | | | | | | | | | |
| + S1BD | –1 | | | | | 1 | | | 1 | | | | | | | 1 | | | 1 | | | | | | | | |
| + S2BD | | –1 | | | | | 1 | | | 1 | | | | | | | 1 | | | 1 | | | | | | | |
| + S3BD | | | –1 | | | | | 1 | | | | | | | | | | 1 | | | | | | | | | |
| + S4BD | | | | –1 | | | | | √ | √ | | | | | | | | | √ | √ | | | | | | | |
| + S1E | –1 | | | | | | | | | | | | | | | | | | | | 1 | | | 1 | | | |
| + S2E | | –1 | | | | | | | | | | | | | | | | | | | | 1 | | | 1 | | |
| + S3E | | | –1 | | | | | | | | | | | | | | | | | | | | 1 | | | | |
| + S4E | | | | –1 | | | | | | | | | | | | | | | | | | | | \ | √ | | |
| – RBL | | | | | | √ | √ | X | √ | √ | | | | | | | | | | | | | | | | | Y |
| + PBL | | | | | | | | X | | | | | | | | | | | | | | | | | | | Y |
| – RCL | | | | | | | | | | | \ | √ | X | √ | √ | | | | | | | | | | | | Y |
| + PCL | | | | | | | | | | | | | X | | | | | | | | | | | | | | Y |
| – RDL | | | | | | | | | | | | | | | | √ | √ | X | \ | √ | | | | | | | Y |
| + PDL | | | | | | | | | | | | | | | | | | X | | | | | | | | | Y |
| – REL | | | | | 1 | | | | | | | | | | | | | | | | \ | \ | X | \ | \ | | Y |
| + PEL | | | | | | | | | | | | | | | | | | | | | | | X | | | | Y |
| + APT | –1 | –A | | | | | | | | | | | | | | | | | | | | | | | | 1 | |
| + PTRQ | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | Y |
| + L1AB | | | | | | \ | \ | | | | | | | | | | | | | | | | | | | | |
| + L2AB | | | | | | \ | \ | | | | | | | | | | | | | | | | | | | | |
| + L1FB | | | | | | | | | \ | √ | | | | | | | | | | | | | | | | | |
| + L2FB | | | | | | | | | \ | √ | | | | | | | | | | | | | | | | | |
| + L1AC | | | | | | | | | | | √ | \ | | | | | | | | | | | | | | | |
| + L2AC | | | | | | | | | | | √ | √ | | | | | | | | | | | | | | | |
| + L1FC | | | | | | | | | | | | | | √ | \ | | | | | | | | | | | | |
| + L2FC | | | | | | | | | | | | | | √ | √ | | | | | | | | | | | | |
| + L1AD | | | | | | | | | | | | | | | | \ | √ | | | | | | | | | | |
| + L2AD | | | | | | | | | | | | | | | | √ | √ | | | | | | | | | | |
| + L1FD | | | | | | | | | | | | | | | | | | | \ | √ | | | | | | | |
| + L2FD | | | | | | | | | | | | | | | | | | | \ | √ | | | | | | | |
| + L1AE | | | | | | | | | | | | | | | | | | | | | \ | \ | | | | | |
| + L2AE | | | | | | | | | | | | | | | | | | | | | \ | \ | | | | | |
| + L1FE | | | | | | | | | | | | | | | | | | | | | | | | \ | \ | | |
| + L2FE | | | | | | | | | | | | | | | | | | | | | | | | \ | \ | | |

FIGURE 6. Linear programming format for Example 1, piecewise linear model.

## 4. THE BRANCH AND BOUND ALGORITHM

We have shown in the previous section that for $r$ routes there are $3^r$ linear programming problems to consider in order to find the optimal solution to the piecewise linear model. This is necessary because there are three cases for each of the $r$ routes. Alternatively, we may consider two mix ratios, $N_{1ij}/N_{2ij}$ and $N_{2ij}/N_{1ij}$, for each route. Each of these may be restricted to be bounded either above or below by its appropriate $L$ limit (see Figure 5).[1] This leads to $2^{2r}=4^r$ cases, and consideration of $4^r$ linear programming problems. This is not at odds with the figure of $3^r$ mentioned above because the two mix ratios on each route are not independent. For each route, therefore, only 3 of the 4 possibilities are realizable. Consideration of the $4^r$ problems facilitates development of the branch and bound algorithm; however, and it does so with no loss in generality. The algorithm, as explained in the following paragraphs, assumes that $L_{1jk}L_{2jk} > 1$ for each route $jk$. Minor modifications would be required if $L_{1jk}L_{2jk} \leq 1$ for some routes $jk$. The branch and bound algorithm may be briefly summarized as follows:

> Consider subsets of the $4^r$ problems and obtain a lower bound on the optimal value in each subset. Partition these subsets into smaller subsets, obtaining newer (and higher) lower bounds, until a subset is found that contains exactly one of the $4^r$ problems and has a solution value that is less than or equal to the lower bound for every other subset. The solution to this problem solves the original problem.

In the following paragraphs we shall characterize the subsets referred to above, describe how they are partitioned into smaller subsets and how the lower bounds are obtained, and specify the order in which subsets are considered. We suggest that these paragraphs be read in conjunction with the example given in Section 5. The algorithm is summarized in the flow diagram shown in Figure 7. Surveys of branch and bound methods are found in [1] and [6].

### Subsets of Problems

A general subset of the $4^r$ possible problems is one determined by specifying that certain of the $2r$ ratios, $n_1$ of them, are restricted to be less than or equal to their $L$ limits, certain ones, $n_2$ of them, are restricted to be greater than or equal to their $L$ limits, and the remaining $2r-n_1-n_2$ ratios are not restricted. This subset thus consists of $2^{2r-n_1-n_2}$ of the $4^r$ problems.

We shall use the following notation to denote subsets of the above type. Let $\sigma$ be a $2r$-dimensional vector representing the above general subset; we shall also refer to $\sigma$ as the subset. The $i^{\text{th}}$ element of $\sigma$ is 0 if the $i^{\text{th}}$ ratio is unrestricted, 1 if it is bounded above by $L_i$, and 2 if it is bounded below by $L_i$.

The general subset $\sigma$ will be partitioned into two different subsets by considering the two values 1 and 2 for a particular element of $\sigma$ that is presently zero, i.e., for a ratio that is presently unrestricted we will consider the two cases of bounding it above and below.

### Lower Bounds

A lower bound on the solution value in the general subset $\sigma$, denoted by $Z(\sigma)$, is the optimal value of the linear programming problem in which the appropriate ratios are bounded above and below (the appropriate $a$, $b$, and/or $c$ values must be used, that is, the appropriate combinations of pieces ①, ②, ③, and ④ in Figure 5), and the remaining ratios are unrestricted (the $a$ and $c$ values corresponding to the increasing portion of the piecewise linear productivity functions are used, that is, pieces ① and ③ of Figure 5). The optimal value of this problem is a lower bound to the solution value in the subset $\sigma$ be-

[1] These bounds replace the mix ratio constraints given by (9) and (10) for case (i), (14) for case (ii), and (18) for case (iii).

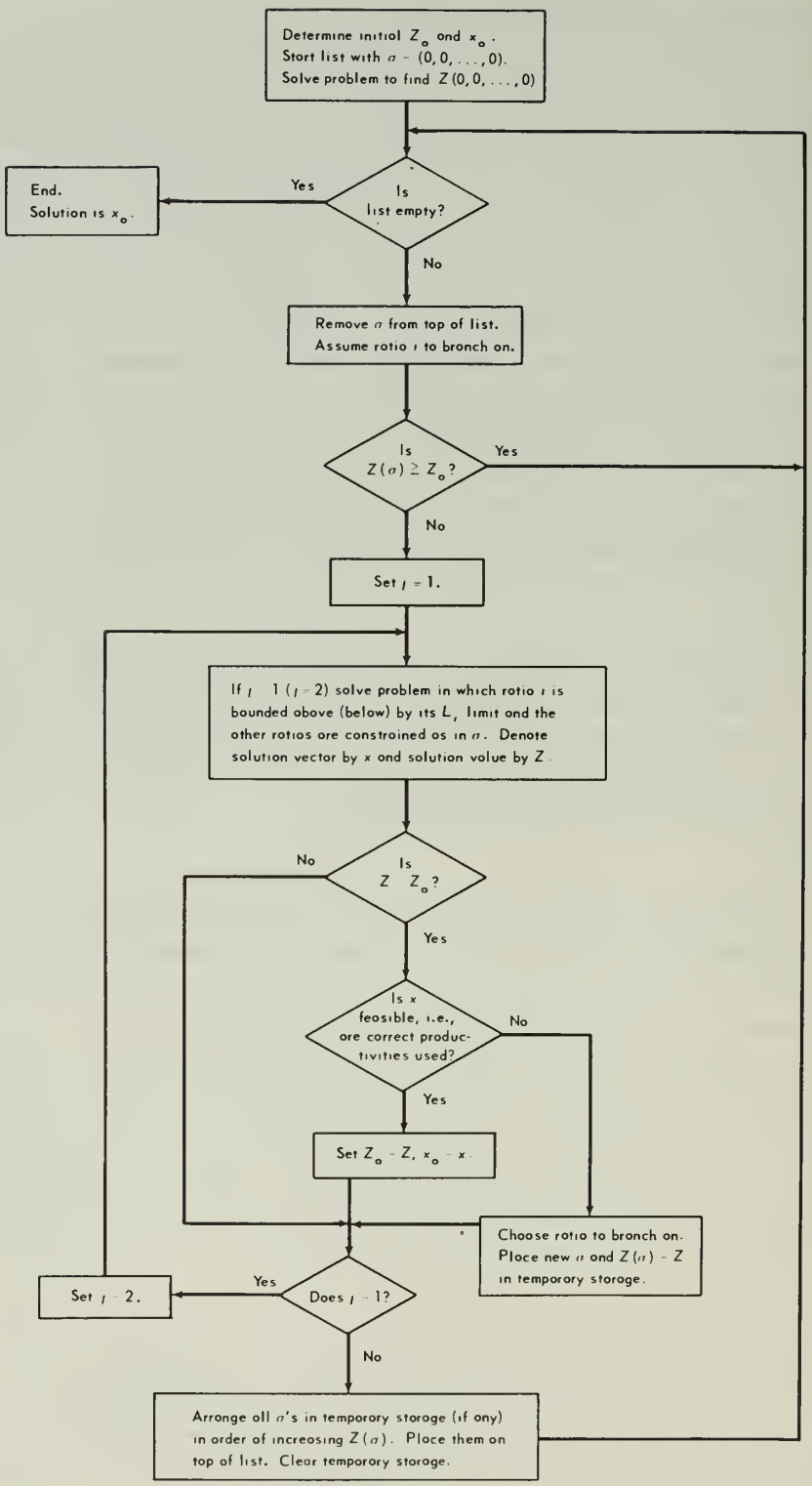



FIGURE 7. Flow diagram of the algorithm.

cause the $a$ and $c$ values corresponding to the increasing portion of the productivity functions are used for all unrestricted ratios. This allows greater productivities than the actual ones if some unrestricted ratios are greater than their $L$ limits. Fewer aircraft will then be required, and the overall cost will be less than the cost obtained by using the actual piecewise linear productivity functions. Let $x(\sigma)$ be the solution vector to the above problem and let $x_0$ be an optimal solution vector.

**Sequential Consideration of Subsets**

The method of determining candidate subsets of the $4^r$ possible problems is best explained in terms of a list of subsets to be considered. Each cycle of the algorithm begins with the consideration of the topmost subset $\sigma$ on the list. This $\sigma$ is partitioned into two smaller subsets, say $\sigma_1$ and $\sigma_2$, and the lower bounds $Z(\sigma_1)$ and $Z(\sigma_2)$ determined as explained above. These two $\sigma$'s are then placed on top of the list with the $\sigma$ having smaller $Z(\sigma)$ on top. The cycle then repeats.

During the operation of the algorithm, the best solution value found to the original problem (denoted by $Z_0$) is kept on record. No $\sigma$ is placed on the list or partitioned into smaller subsets unless $Z(\sigma) < Z_0$. If a solution to a problem uses the correct productivities for all ratios, this solution is feasible for the original problem and if the solution value is less than $Z_0$ this value becomes the new value of $Z_0$. An initial value of $Z_0$ is found by solving the problem in which all $2r$ ratios are bounded above by their $L$ limits. This choice of one of the $3^r$ realizable problems is based on the belief that use of the increasing portions of the productivity curves with moderate values of all the ratios will often produce a good value of $Z_0$. Any other problem, or none at all (in which case $Z_0 = \infty$ initially), could be substituted.

We must specify how the algorithm chooses a ratio to "branch" upon, i.e., how a particular zero element of $\sigma$ is chosen as the basis of partition of $\sigma$ into $\sigma_1$ and $\sigma_2$. This is done by examining the solution vector $x(\sigma)$ to the problem yielding $Z(\sigma)$ as its solution value. Each ratio is computed and compared with its $L$ limit, and the one that is the largest number of times its $L$ limit is chosen. In case of a tie (frequently several ratios are infinite) the ratio chosen is that one associated with the greatest amount of materiel transported over and above what could be transported with the correct productivities for the given values of $N_{1jk}$ and $N_{2jk}$.

## 5. AN ILLUSTRATIVE EXAMPLE

We shall illustrate the branch and bound algorithm with a very small hypothetical example. The deployment network is shown in Figure 8; country $A$ must supply countries $B$, $C$, $D$, and $E$ using only type 1 and type 2 aircraft. There are two contingencies: (1) countries $B$, $C$, and $E$ must be supplied simultaneously, and (2) countries $B$, $D$, and $E$ must be supplied simultaneously. Figure 9 shows the linear programming format for this example. The notation is consistent with that used in Figure 6 for Example 1. Figure 10 gives the parameters of the piecewise linear productivity functions for the two types of planes on the four routes. Note that the fractional change in productivity from minimum to maximum

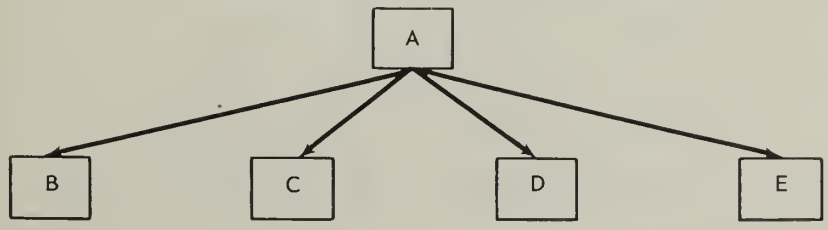



FIGURE 8. Deployment network for Example 2.

| Constraint name | Variable | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | S 1 | S 2 | N 1 A B | N 2 A B | N 1 A C | N 2 A C | N 1 A D | N 2 A D | N 1 A E | N 2 A E | R H S |
| Cost | 5 | 2.4 | | | | | | | | | |
| + S1L | 1 | | | | | | | | | | 350 |
| + S2L | | 1 | | | | | | | | | 1500 |
| + S1BCE | –1 | | 1 | | 1 | | | | 1 | | |
| + S2BCE | | –1 | | 1 | | 1 | | | | 1 | |
| + S1BDE | –1 | | 1 | | | | 1 | | 1 | | |
| + S2BDE | | –1 | | 1 | | | | 1 | | 1 | |
| – RBL | | | √ | √ | | | | | | | 135 |
| – RCL | | | | | √ | √ | | | | | 135 |
| – RDL | | | | | | | √ | √ | | | 135 |
| – REL | | | | | | | | | √ | √ | 135 |
| + L1AB | | | √ | √ | | | | | | | |
| + L2AB | | | √ | √ | | | | | | | |
| + L1AC | | | | | √ | √ | | | | | |
| + L2AC | | | | | √ | √ | | | | | |
| + L1AD | | | | | | | √ | √ | | | |
| + L2AD | | | | | | | √ | √ | | | |
| + L1AE | | | | | | | | | √ | √ | |
| + L2AE | | | | | | | | | √ | √ | |

FIGURE 9. Linear programming format for Example 2, piecewise linear model.

| Productivity | a | b | L | Ratio | Ratio Designation |
|---|---|---|---|---|---|
| $P_{1AB}$ | 1.000 | 1.100 | 5 | $N_{2AB}/N_{1AB}$ | $r_2$ |
| $P_{1AC}$ | 1.000 | 2.000 | 5 | $N_{2AC}/N_{1AC}$ | $r_4$ |
| $P_{1AD}$ | 1.000 | 1.100 | 5 | $N_{2AD}/N_{1AD}$ | $r_6$ |
| $P_{1AE}$ | 1.000 | 2.000 | 5 | $N_{2AE}/N_{1AE}$ | $r_8$ |
| $P_{2AB}$ | 0.250 | 0.275 | 2 | $N_{1AB}/N_{2AB}$ | $r_1$ |
| $P_{2AC}$ | 0.250 | 0.275 | 2 | $N_{1AC}/N_{2AC}$ | $r_3$ |
| $P_{2AD}$ | 0.250 | 0.500 | 2 | $N_{1AD}/N_{2AD}$ | $r_5$ |
| $P_{2AE}$ | 0.250 | 0.500 | 2 | $N_{1AE}/N_{2AE}$ | $r_7$ |

FIGURE 10. Productivity parameters and ratio designations for Example 2.

is considerably different for the different routes and airplane types. Also shown in Figure 10 are the ratios determining the productivities and the positions assigned the ratios in the $\sigma$ vectors.

A tree illustrating the branch and bound algorithm is shown in Figure 11. Each node corresponds to a linear programming problem that was solved, and the numerical sequence of the nodes indicates the sequence in which the problems were solved. Shown at each node are the vector $\sigma$ and the lower bound $Z$ found at that node. At node 1 the initial value of $Z_0$ is also shown; this was obtained by solving the problem in which all ratios were bounded above by their $L$ limits. Subsequent values of $Z_0$ were found at nodes 8 and 9, as is indicated on the tree. Also shown is the ratio that was branched on in each case. One can see, for example, that the optimal solution found at node 9 has $r_1 \geq L_1$, $r_7 \leq L_7$, $r_3 \geq L_3$, and $r_5 \geq L_5$.

The values of the variables at the optimal solution are shown in Figure 12. In this problem it turns out that node 11 would also lead to the optimal solution found at node 9. This is because $r_7 = L_7$ at the optimal solution, and it illustrates the fact that the path through the tree to an optimal solution is frequently not unique.

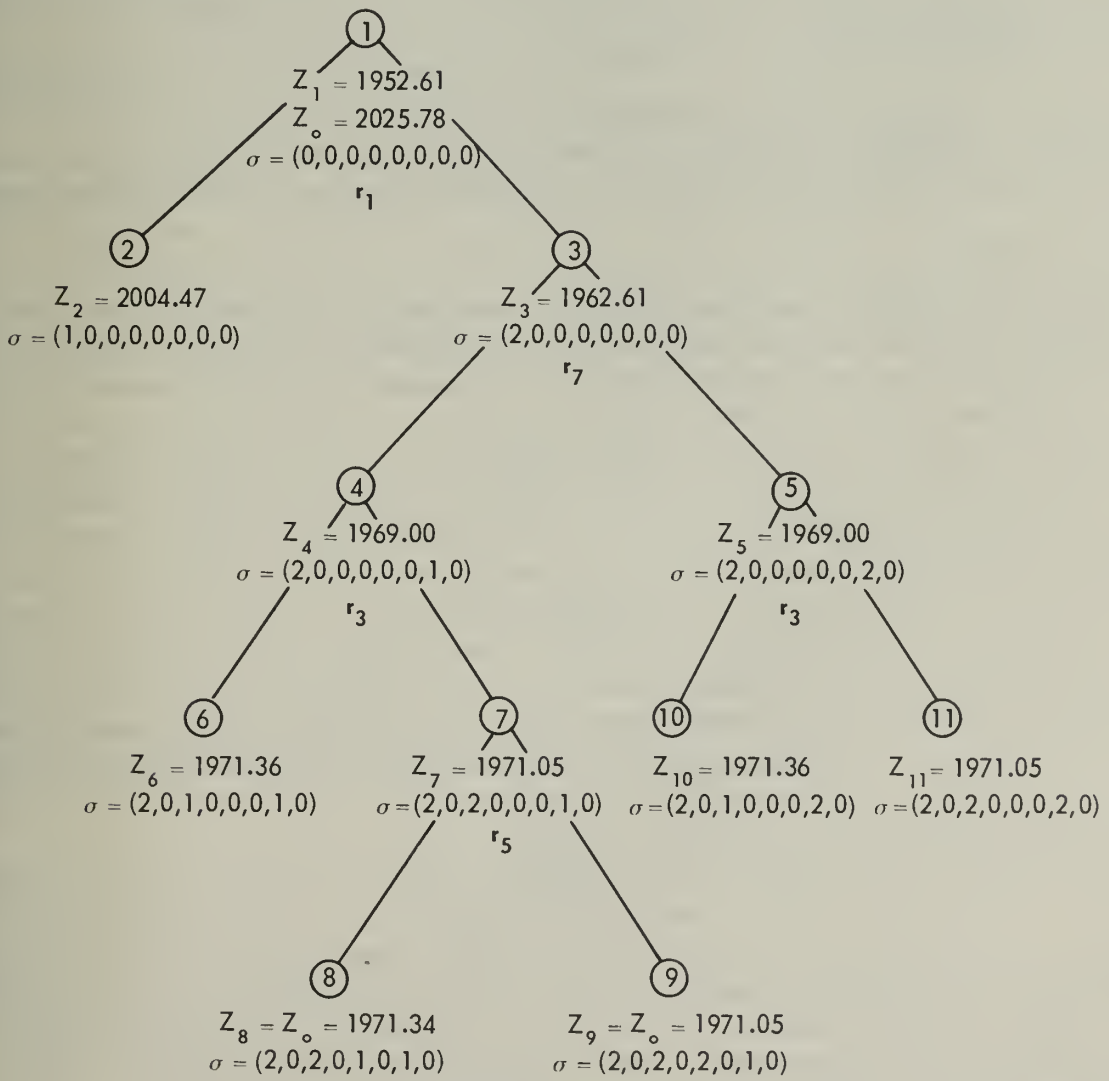



FIGURE 11. Tree for Example 2.

Cost = 1971.05
$S_1$ = 350.00
$S_2$ = 92.10

| | | | |
|---|---|---|---|
| $N_{1AB}$ = | 135.00 | $N_{1AD}$ = | 113.10 |
| $N_{2AB}$ = | 0.00 | $N_{2AD}$ = | 42.10 |
| $N_{1AC}$ = | 115.00 | $N_{1AE}$ = | 100.00 |
| $N_{2AC}$ = | 42.10 | $N_{2AE}$ = | 50.00 |

FIGURE 12. Optimal solution for Example 2.

It is of interest to compare the optimal value for this example, 1971.05, with the optimal value obtained if it is assumed that all productivities are constant and equal to the numbers shown in the "a" column of Figure 10. This value is 2278.00, 15.5 percent above the previous figure. This is for a small hypothetical problem; however, and may not be indicative of the savings in general. This point will be discussed further in Section 6.

The algorithm was required to solve 11 linear programming problems, 12 counting the one used to obtain the initial value of $Z_0$, in order to solve Example 2. This is about 15 percent of $3^r = 3^4 = 81$. For Example 1, described in Section 1, the algorithm was required to solve 199 problems, approximately 3 percent of $3^8 = 6561$. The significance of these numbers will be discussed in the next section.

## 6. DISCUSSION

Based upon the limited experience reported in the previous section, it appears that our algorithm is less efficient than one developed in [5] for a somewhat different class of problems. The latter algorithm also solves a nonconvex programming problem by solving a small fraction of a number of possible convex problems, but it is for problems in which only the objective function is nonconvex whereas our problem has nonconvex constraints. We can identify two sources of this apparent inefficiency: (1) The type of strategic deployment model considered here usually has some redundant constraints that do not affect the minimum cost, but do prevent the algorithm from terminating. For example, in Figure 11 nodes 7 and 9 have the same optimal cost, but the solution at node 7 is not feasible because the ratio $N_{1AD}/N_{2AD}$ is greater than $L_{2AD}$. At node 9 the productivity $P_{2AD}$ is reduced, but the cost remains the same because constraint *S1BDE* was redundant at node 7 (the value of $S_1$ required to satisfy constraint *S1BCE* was more than sufficient to satisfy constraint *S1BDE)* so that $N_{1AD}$ could be increased to compensate for the lower productivity of type 2 planes on route *AD* without increasing the cost. (2) When some ratios are equal to their $L$ limits at an optimal solution there are multiple paths through the tree to that solution, and the algorithm may have to trace out a large fraction of these or portions of them. This was observed in connection with Example 2.

The efficiency of the algorithm can be increased by making comparisons in the flow diagram of Figure 7 with $Z_0(1-\rho)^{-1}$ or $Z_0 - \eta$ instead of with $Z_0$. Here $\rho$ and $\eta$ are appropriate positive constants. Comparison with the smaller value will tend to rule out more subsets $\sigma$ and thus speed the algorithm. The algorithm then guarantees an optimal value to within a fraction $\rho$ in the first case and to within an absolute amount $\eta$ in the second case. In Examples 1 and 2, we set $\rho$ to $4 \times 10^{-6}$. A value of $\rho$ of $10^{-3}$ would reduce the number of problems solved from 199 to 51 for Example 1, but would not reduce the number solved for Example 2.

It is a tenet of operations research that the simplest model consistent with reality be used. In some problems the change in productivity as the mix ratio changes may be small enough that the overall cost obtained from the model using conservative constant productivities is only slightly more than the cost obtained from the piecewise linear model. In view of the uncertainty associated with the values of coefficients appearing in such a problem, this difference in cost might not be significant. It would not therefore be advantageous to use the piecewise linear model in such a case, especially since the computation time of the branch and bound algorithm, although extremely small relative to the time required to exhaust the $3^r$ possibilities, is still many times the time required to solve a linear programming problem with constant productivities. There may be situations, other than deployment problems, however, where use of the more complex piecewise linear model can be justified. One case in which productivity might be heavily dependent on aircraft mix is that in which the two aircraft types are bombers and fighters and productivity is measured in damage to the enemy per aircraft. The ratio of fighters to bombers on a particular mission could significantly affect both the damage done per bomber and the rate of survival of the bombers. Generalizing further, a problem of selecting the optimal weapon system mix when the unit effectiveness of a particular weapon system is not constant but instead depends on the ultimate mix selected could be solved by the methodology developed in this paper.

It is possible to extend the model and algorithm presented here by using piecewise linear approximations to the productivity functions with more than two linear pieces. This would provide a more accurate approximation of the productivity functions and would not significantly complicate the branch and bound algorithm. The number of possible linear programming problems would increase from $3^r$; however, and the number of problems solved by the algorithm would also increase.

The model could also be extended by assuming that productivities are dependent upon several vehicle ratios instead of one. Specifically, let $P_i$ be the productivity of vehicle type $i$ on a particluar route and let $N_j (j=1, \ldots, p)$ be the number of vehicles of type $j$ on that route. Assume that

$$P_i = \sum_{j=1}^{p} c_{ij}(N_j/N_i),$$

so that the amount delivered by vehicle type $i$ is

$$P_i N_i = \sum_{j=1}^{p} c_{ij} N_j.$$

Here $c_{ij}$ is a different constant for different ranges of the ratio $(N_j/N_i)$, just as in the model developed in Section 3. Actually Example 1 of Sections 2 and 3 considered three vehicle types, two types of aircraft plus ships, but no interaction between planes and ships was assumed.

## REFERENCES

[1] Agin, N., "Optimum Seeking with Branch and Bound," Management Science, Series B, *13*, 176–185 (December 1966).

[2] Fiacco, A. V. and G. P. McCormick, "The Sequential Unconstrained Minimization Technique for Nonlinear Programming: A Primal-Dual Method," Management Science, *10*, 360–366 (1964).

[3] Fitzpatrick, G. R., J. Bracken, M. J. O'Brien, L. G. Wentling and J. C. Whiton, "Programming the Procurement of Airlift and Sealift Forces: A Linear Programming Model for Analysis of the Least-Cost Mix of Strategic Deployment Systems," Naval Research Logistics Quarterly, *14*, 241–255 (June 1967).

[4] Gross, D., "The Effects of Aircraft Mix on Aircraft Productivities in a Least Cost, Strategic Deployment Linear Programming Model," Working Paper, Research Analysis Corporation, McLean, Virginia (September 15, 1967).

[5] Jones, A. P. and R. M. Soland, "A Branch-and-Bound Algorithm for Multilevel Fixed-Cost Problems," RAC–TP–285, Research Analysis Corporation, McLean, Virginia (October 1967).

[6] Lawler, E. L. and D. E. Wood, "Branch-and-Bound Methods: A Survey," Operations Research, *14*, 699–719 (1966).

[7] Regan, L. G., W. E. Billion, R. D. Goodfriend, R. Melby and L. A. Poth, "The Peacetime Economic Value of Intertheater Lift Systems," Interim Draft Report, Research Analysis Corporation, McLean, Virginia (February 1968).

[8] Whiton, J. C., "Some Comments on Suboptimization," Working Paper, Research Analysis Corporation, McLean, Virginia (March 17, 1967).

* * *

# MARKOV CHAIN ANALYSES OF MULTIPROGRAMMED COMPUTER SYSTEMS [1]

E. G. Coffman, Jr.

*Princeton University*

**ABSTRACT**

Most operating systems for large computing facilities involve service disciplines which base, to some extent, the sequencing of object program executions on the amount of running time they require. It is the object of this paper to study mathematical models of such service disciplines applicable to both batch and time-shared processing systems. In particular, Markov queueing models are defined and analyzed for round-robin and foreground-background service disciplines. With the round-robin discipline, the service facility processes each program or job for a maximum of $q$ seconds; if the program's service is completed during this quantum, it leaves the system, otherwise it returns to the end of the waiting line to await another quantum of service. With the foreground-background discipline each new arrival joins the end of the foreground queue and awaits a single quantum of service. If it requires more it is subsequently placed at the end of the background queue which is allocated service only when the foreground queue is empty.

The analysis focuses on the efficiency of the above systems by assuming a swap or set-up time (overhead cost) associated with the switching of programs on and off the processor. The analysis leads to generating functions for the equilibrium queue length probabilities, the moments of this latter distribution, and measures of mean waiting times. The paper concludes with a discussion of the results along with several examples.

## I. INTRODUCTION

This paper concerns two particular models of computing systems in which the sequencing of program (or job) operations is based on running-time priorities; in subsequent sections these models are mathematically defined, analyzed, and their performance studied. The computing systems studied can be considered to fall within two major categories; batch-processing systems and time-sharing systems. As we shall see, such a categorization is not sufficient to encompass all systems of interest; the service disciplines we shall study may well apply to systems (e.g., certain real-time processing systems) that must be considered as combinations of the above types of systems.

Running-time priority disciplines are those which discriminate between programs on the basis of the amount of service (running-time) they require. Discrimination is made beforehand in the case of conventional priority queues [19] and in the case of the model analyzed by Phipps [12], in which the priority rule is shortest-job first. In the models of particular interest here, however, the discrimination can be made only after programs are partially run, since it is assumed that neither the exact running times nor any indications of the running times of arriving programs are known in advance. Precisely how this can be done will be clear below when we describe in detail the queueing models to be analyzed in the following sections.

[1] Preparation of this paper was sponsored in part by the Office of Naval Research and the U.S. Atomic Energy Commission.

### The Round-Robin (RR) Model

Queueing systems in which the service discipline is the so-called round-robin discipline have been the object of considerable analysis [1, 5, 8, 11, 14, 15] in the past few years, primarily in connection with time-sharing applications. As shown in Figure 1, programs (or jobs) arrive to the queue

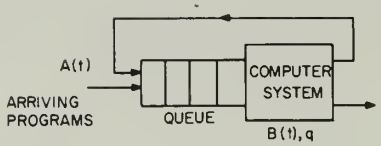



FIGURE 1. The Round-Robin Model.

of programs awaiting service on the computer system. The stochastic input process is described by an interarrival time distribution, identical and independent for each interarrival period. We shall denote this distribution by $A(t)$.

The service time of all programs arriving to the system is subject to the same stationary probability distribution which we shall denote by $B(t)$. Programs are taken from the queue first-come-first-served and provided with a certain fixed amount of running time which we shall call a quantum ($q$). If the program being operated completes within the time interval allocated then it is simply ejected from the system. If, on the other hand, it requires more time to complete then it is removed from the computer and put back to the end of the line. In due course, after the other programs in line ahead of this program have received their quantum of service, the interrupted program is again operated, continuing from the point at which the previous operation was interrupted; i.e., we have a "preemptive resume" rule implying that running time was not lost because of interruption. The procedure as outlined is continued for all programs in the queue; each program makes as many of the "loops," shown in Figure 1, as needed to complete its total running time requirement.

### The Foreground-Background (FB) Model

This model, which arises in connection with both batch-processing and time-sharing systems, is shown in Figure 2. As can be seen, the input mechanism and the service time distribution are the

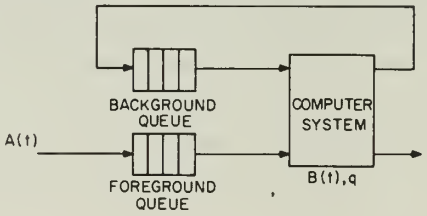



FIGURE 2. The Foreground-Background Model.

same as for the RR model in Figure 1. In this model we shall assume that a program arriving at an empty system is operated to completion; however, programs arriving to a non-empty system are again allocated some limited quantum ($q$) of service when they first arrive at the service point. If a program completes within this time it leaves the system as before. If it requires additional time, however, it is placed at the end of a second queue consisting of other, previous programs that also required more than the quantum allocated to them. Now the second queue is only serviced when there are no new arrivals in the system waiting or being worked on (i.e., when the first queue is empty and no new arrival is being serviced). For reasons that should be clear, the new arrival queue is called the *foreground* queue while the queue of programs having already received a quantum of service is called the *back-*

*ground* queue. Although the FB model has two queues it is important to note that only the foreground queue is supplied from the input source. (We view arrivals to an empty system as being immediately switched from the foreground queue to the background queue.) In general, programs serviced from the background queue will also be allocated time on a quantum basis; however, depending on the application, the background quantum will usually be larger than the foreground quantum and may even be infinite (in which case, background programs are served to completion).

**Costs Associated with the RR and FB Service Disciplines**

The RR and FB models just described were first discussed in connection with time-sharing systems [3, 9, 16]. In both types of systems the principal objective of the quantum-controlled service discipline is to provide a faster operation or turn-around time for those programs or jobs with small computational requirements as compared with the turn-around time they would receive with a conventional first-come-first-served, run-to-completion discipline. The costs incurred in obtaining this type of performance are (1) the waiting times of the longer running programs, and (2) the loss in operating efficiency brought about by the switching of programs and the extra computer time taken up by the system executive. These latter "overhead" costs will be represented in the RR and FB models by associating a cost parameter $\tau$ with every switching of programs that takes place in the execution of the respective service disciplines. In current computing systems the cost $\tau$ is predominantly that of swap time; hereafter, we shall ignore the operation time taken by the executive and refer to $\tau$ as the swap time. In a completely general analysis, $\tau$ should be considered a random variable; however, we shall assume initially that $\tau$ is a constant. Later on we shall indicate the way in which a random $\tau$ can be handled in the analysis.

**Description of the Analysis**

Regarded as priority models, the models we have presented are distinguished by the fact that priorities are not known beforehand, but are based on dynamic operating behavior. Of course, this distinction is fundamental, and as a result the treatments of priority models that exist in the queueing theory literature can be carried over only in a limited way. Despite structural differences in the models, analyses similar to the type we shall be using in this paper can be found, for example, in the work of Krishnamoorthi and Wood [8], Miller [10], and Coffman [1]. In particular, our analysis will be similarly based on the theory of imbedded Markov chains as developed by Kendall [4].

Efficient computer operation requires that the sequencing of computer operations be controlled automatically by a control program within an operating or executive system designed for this purpose. Thus, the mechanism for instituting a service discipline ordinarily observes the state of the system only at instants between the execution of consecutive programs. Thus, the results we shall obtain for a Markov chain defined at these instants will coincide with system behavior as observed by the control program making scheduling decisions. Of course, as is the case with most analyses of this kind, the results are not generally applicable for arbitrary instants in (continuous) time. Depending on the particular case, however, these latter results when desired are closely approximated by the results in discrete time. Further studies of computer operation under running time priority service disciplines can be found in references [2, 6, 17] of the bibliography.

## II. THE ANALYSIS

### The Round-Robin Model

The first model to be analyzed will be the round-robin model of Figure 1 first introduced by Kleinrock [5]. Our principal objective will be an analysis from which we may determine the effects of swapping and overhead on the mean number in the system and the mean waiting time. We shall pass from the discrete, geometrically distributed service and interarrival times assumed by Kleinrock, to continuous, exponentially distributed times with parameters $\mu$ and $\lambda$, respectively. As is well-known, these distributions, the exponential and geometric, are analogous in the sense that they both possess the memoryless property; the former in continuous time and the latter in discrete time. Also, in contrast to the finite source models of Coffman, Krishnamoorthi, and Wood [2, 8] we shall consider the infinite source assumption. As a consequence we shall obtain somewhat simpler results.

The basic random process of interest in this model is the number of jobs in the system as a function of time; we denote this process by $\xi(t)$. From the definition of the round-robin service discipline it follows that the quantum constraint on program operation is such that $\xi(t)$ is definitely not a Markov process; however, we shall consider a "chain" (i.e., a sequence $\xi(t_k)$ defined on discrete time points $t_k$ of $\xi(t)$, $k=1, 2, 3, \ldots$) imbedded in the original process in such a way that the Markov property holds for the sequence $\xi(t_k)$. The equilibrium probability distribution for $\xi(t_k)$ will then be computed using classical methods. The advantage of this approach over other approaches to our problem is the greatly reduced complexity in the results we seek.

For the epochs $t_k$ of our Markov chain, there exist several possible choices which differ mainly in their definition during idle periods (i.e., when the service facility is idle). Following Krishnamoorthi and Wood [8], we shall select for the $t_k$ the instants just after a job completion or a quantum expiration, whichever occurs first. Accordingly, the distribution of the time interval $(t_{k+1}-t_k)$ between successive epochs will be as follows for inter-epoch intervals not commencing with an idle period.

$$F(x)=Pr[(t_{k+1}-t_k)<x]=\begin{cases}0; & x<\tau\\ 1-\epsilon^{-\mu(x-\tau)}; & \tau\leq x<q+\tau,\\ 1; & x\geq q+\tau\end{cases} \tag{1}$$

where $q$ is the quantum size and $\tau$ is the swap time (any other overhead costs are assumed to be lumped into the swap time parameter). We denote by $m_s$ the mean of this distribution. We have

$$m_s=(1/\mu)\ (1-\epsilon^{-\mu q})+\tau. \tag{2}$$

Due to the memoryless property of the exponential distribution, the amount of service required by a program at the service facility has the same exponential distribution (with mean $1/\mu$ sec) regardless of how many quanta of service the given program had received previously. Thus, with a Poisson input process it is clear that $\xi(t_k)$ does indeed constitute a homogeneous Markov chain, and will be completely described by the one-step transition probabilities which we now proceed to write out. Let $p_{ij}$ be the probability that there are $j$ programs in the system at the time $t_{k+1}$ given that there were $i$ programs at time $t_k$. In short

$$p_{ij}=Px[\xi(t_{k+1})=j|\xi(t_k)=i] \qquad k=1, 2, 3, \ldots.$$

Then, for all $j<i-1$ we have $p_{ij}=0$. For $j=i-1\geq 0$ we have,

$p_{ij}=$ Pr [0 arrivals in $t+\tau$ secs; program in service completes in $t\geq q$ secs] and for $j>i-1$ we have;

$p_{ij}=Pr[(j-i)$ arrivals in $q+\tau$ secs; program in service does not complete in $q$ secs)]
$+Pr[(j-i+1)$ arrivals in $t+\tau$ secs; program completes $t<q$ secs].

Finally, we have $p_{0j}=p_{1j}$ for all $j\geq 0$. This last relation follows from the observation that in order to pass from 0 programs to $j=0$ programs, we must always service for one quantum the first arrival. But since the idle period has the memoryless exponential distribution, we may, for the purposes of calculating $p_{0j}$, replace $t_k$ by the time instant immediately following the time of the first arrival. If we let $P(n|t)$ denote the probability that in time $t$ there are $n$ arrivals from an infinite Poisson source with average rate $\lambda$, then we may use the definition of the service time distribution to rewrite the above expression for the $p_{ij}$ as follows:

(3a)
$$p_{ij}=\begin{cases} 0; & j<i-1 \\ \int_0^q P(0|t+\tau)\mu\epsilon^{-\mu t}dt; & j=i-1\geq 0 \\ P(j-i|q+\tau)\epsilon^{-\mu q}+\int_0^q P(j-i+1|t+\tau)\mu\epsilon^{-\mu t}dt; & j\geq i\geq 1 \end{cases}$$

(3b)

(3c)

(3d)
$$p_{0j}=p_{1j};\ j\geq 0.$$

A limiting, stationary probability distribution

$$\{\pi_j\}_{j=0}^{\infty}$$

for the number of programs in the system at the epochs of the Markov chain $\xi(t_k)$ will exist if the average input rate does not exceed the maximum output rate of programs [18]. Under this condition we have

(4)
$$\pi_j=\sum_{i=0}^{\infty} p_{ij}\pi_i.$$

We now set about finding for the probabilities $\pi_j$ a generating function

(5)
$$T(z)=\sum_{j=0}^{\infty}\pi_j z^j$$

(where $z$ is any number lying within or on the unit circle) such that $T(z)$ is expressed in terms of the parameters $\lambda$, $\mu$, $\tau$, $q$. Multiplying Eq. (4) by $z^j$ and summing from zero to infinity yields

$$T(z)=\sum_{j=0}^{\infty} z^j\sum_{i=0}^{\infty} p_{ij}\pi_i.$$

Reversing the above summations and separating the result into the different cases represented in Eq. (3) give

(6)
$$T(z)=\pi_0\sum_{j=0}^{\infty} p_{0j}z^j+\sum_{i=1}^{\infty}\pi_i\left\{p_{i,i-1}z^{i-1}+\sum_{j=i}^{\infty} p_{ij}z^j\right\}.$$

Let

$$P(z)=\sum_{n=0}^{\infty} z^n\int_0^q P(n|t+\tau)\mu\epsilon^{-\mu t}dt$$

(7)
$$Q(z)=\sum_{n=0}^{\infty} z^n\epsilon^{-\mu q}P(n|q+\tau).$$

By using Eqs. (3) and (7), Eq. (6) may be rendered as

$$T(z)=\pi_0\,[P(z)+zQ(z)]+[T(z)-\pi_0]\left[\frac{P(z)}{z}+Q(z)\right],$$

from which

$$T(z)=\frac{(1-z)}{1-z\beta(z)}\,\pi_0, \tag{8}$$

where

$$\beta(z)=[P(z)+zQ(z)]^{-1}.$$

It remains to determine $\pi_0$. For this we use the relation

$$\lim_{z\to 1} T(z)=1$$

which is easily derived from Eq. (5). Noting that $\beta(1)=1$, from Eq. (7) it can be seen that both the numerator and denominator of Eq. (8) vanish at $z=1$. Thus, an application of l'Hospital's rule yields,

$$\lim_{z\to 1} T(z)=\lim_{z\to 1}\frac{\pi_0}{\beta(z)+z\beta'(z)}=1.$$

Solving for $\pi_0$ gives

$$\pi_0=1+\beta'(1). \tag{9}$$

Let us now evaluate $P(z)$ and $Q(z)$ in order to work out $P'(1)$, $Q'(1)$, and therefore $\beta'(1)$. From Eq. (7) we have

$$P(z)=\int_0^q \sum_{j=0}^{\infty} \epsilon^{-\lambda(t+\tau)}\,\frac{[\lambda z(t+\tau)]^j}{j!}\,\mu\epsilon^{-\mu t}dt, \tag{10}$$

where we have taken advantage of the fact that within the circle of convergence of the above power series we may reverse the order of summation and integration. Evaluation of Eq. (10) gives

$$P(z)=\frac{\epsilon^{-\lambda\tau(1-z)}}{1+\rho(1-z)}\,\{1-\epsilon^{-\mu q[1+\rho(1-z)]}\}, \tag{11}$$

where $\rho=\lambda/\mu$. From Eq. (11) the following expression is obtained for $P'(1)$:

$$P'(1)=\rho(1+\mu\tau)\,(1-\epsilon^{-\mu q})-\rho\mu q\epsilon^{-\mu q}. \tag{12}$$

For $Q(z)$ we obtain

$$Q(z)=\epsilon^{-\lambda\tau(1-z)}\epsilon^{-\mu q[1+\rho(1-z)]}, \tag{13}$$

from which

$$Q'(1)=\rho(\mu q+\mu\tau)\epsilon^{-\mu q}. \tag{14}$$

Finally, from Eq. (8) we obtain

$$\beta'(1)=-[P'(1)+Q(1)+Q'(1)], \tag{15}$$

which establishes on substitution into Eq. (9)

$$\pi_0 = (1-\rho)\ (1-\epsilon^{-\mu q}) - \lambda\tau. \tag{16}$$

From Eq. (8) the first two moments of the distribution can be calculated using the following relationships:

$$\begin{aligned} \bar{n} &= \lim_{z\to 1} T'\ (z) \\ \mathrm{Var}(n) &= \lim_{z\to 1} [T''(z) + T'(z)(1-T'(z))], \end{aligned} \tag{17}$$

which are easily derived from Eq. (5). In particular, we have for the mean number in the system,

$$\bar{n} = \frac{(1-\pi_0) - \beta''(1)/2}{\pi_0}. \tag{18}$$

To determine $\beta''(1)$ from Eq. (8), we must evaluate $P''(1)$ and $Q''(1)$ from Eqs. (11) and (13). For this we have

$$\beta''(1) = [P''(1) + Q''(1) + 2Q'(1)] - 2[P'(1) + Q'(1) + Q(1)]^2;$$

using Eqs. (12) and (14), we find

$$\beta''(1) = 2\lambda q\delta + 2\delta^2(1-\rho^2) + 2\rho\delta(\rho - \lambda q) - (\lambda\tau + \delta)[\lambda\tau + 2\rho(1-\rho)],$$

where

$$\delta \equiv \epsilon^{-\mu q}.$$

It is easy to verify from the above equation, with the help of Eqs. (16) and (18), that if we allow $\tau \to 0$ and $q \to \infty$ we have $\beta''(1) = 0$ and

$$\pi_0 = 1-\rho, \qquad \bar{n} = \frac{\rho}{1-\rho}, \tag{19}$$

which correspond to the results of Erlang's classical model [13]. The above affords a partial check of our results.

**Conditional Waiting Times for the RR Model**

For the purpose of analyzing the effects of swap time on waiting times we now derive an expression for the mean waiting time in queue of an arrival to the RR system conditioned on the service required. In particular, we shall be concerned with the waiting time of a new arrival from the point of view of the system, as discussed earlier. Alternatively, we may consider the waiting time of a program arriving at the RR system in equilibrium where we assume the time of arrival coincides with a regeneration point of the Markov chain $\xi(t_k)$.

Let the service time required by the above program arrival be denoted by $s$, let $W_s$ be the mean, conditional waiting time *in queue* of this program, and let $m$ be defined as an integer such that for the value of $s$ to which it corresponds we have

$$0 \leq mq - s < q. \tag{20}$$

Then, we have

$$W_s = \frac{m_s}{1-\alpha}\left[m\lambda(q+\tau) + \left(\bar{n} - \frac{\lambda(q+\tau)}{1-\alpha}\right)(1-\alpha^m)\right], \tag{21}$$

where $\lambda$ and $\tau$ retain their prior definitions, $m_s$ is given by Eq. (2), $\bar{n}$ is the average number of programs in the system from Eq. (18), and $\alpha$ is given by

$$\alpha = \delta + \lambda m_s. \tag{22}$$

For the proof that we now give for Eq. (21) we shall employ expected value arguments and a method essentially similar to that used by Kleinrock [5] for the RR model in discrete time. Let $y_i$ denote the time spent in queue on the $i$th pass by the program (the "tagged" program) whose waiting time we are seeking. Clearly,

$$W_s = E\left\{\sum_{i=1}^{m} y_i\right\}. \tag{23}$$

Correspondingly, we define $N_i$ as the mean number of programs ahead of the tagged program at the beginning of the $i$th pass. We shall now develop a general expression for $N_i$. First of all, we note that $N_1 = \bar{n}$ by definition of $\bar{n}$. For $i > 1$, $N_i$ will be made up of the mean number of those programs of $N_{i-1}$ whose processing requirements exceed $q$ secs (we call these returning programs) and the mean number of new arrivals that occur during the time interval $y_{i-1} + q + \tau$. (The $q + \tau$ secs is included because of the tagged program's operation following $y_{i-1}$.) Thus we have;

$$N_i = \delta N_{i-1} + \lambda(\bar{y}_{i-1} + q + \tau); \qquad i > 1. \tag{24}$$

In Eq. (24) $\bar{y}_{i-1}$ will simply be $(q+\tau)$ times the mean number of returning programs plus the mean number of departing programs times their mean time of operation within the quantum. Hence by definition of $m_s$,

$$\bar{y}_{i-1} = N_{i-1} m_s. \tag{25}$$

Substitution of Eq. (25) into Eq. (24) now gives

$$N_i = N_{i-1}[\delta + \lambda m_s] + \lambda(q+\tau)$$

or

$$N_i = \alpha N_{i-1} + \lambda(q+\tau),$$

according to the definition of Eq. (22). Now solving this equation for $N_i$ with the condition $N_1 = \bar{n}$ yields

$$N_i = \alpha^{i-1}\bar{n} + \lambda(q+\tau)\sum_{k=0}^{i-2} \alpha^k; \qquad i > 1. \tag{26}$$

By use of induction, Eq. (26) is easily established. From Eqs. (23) and (25) we may now write

$$W_s = E\left\{\sum_{i=1}^{m} y_i\right\} = m_s \sum_{i=1}^{m} N_i,$$

whereupon substitution of Eq. (26) yields, after carrying out the summations,

$$W_s = \frac{m_s}{1-\alpha}\left[m\lambda(q+\tau) + \left(\bar{n} - \frac{\lambda(q+\tau)}{1-\alpha}\right)(1-\alpha^m)\right].$$

**The Foreground-Background Model**

We shall now analyze, using the same basic theory as before, the foreground-background (FB) model described earlier. The specific Markov chain that we first analyze has the same type of definition

as for $\xi(t_k)$ in the round-robin model. That is, the $t_k$ will again be those instants immediately following a program completion or quantum interruption, whichever occurs first according to the FB discipline. For the present model, however, the state of the system at $t_k$ will be given by a pair of numbers $(m, n)$ representing the number of foreground programs and background programs, respectively, in the system. As a result, the overall effect will be to extend our previous analysis to two dimensions.

According to the earlier description there are three distinct modes of system operation (defined at the epochs $t_k$) that we must consider in order to determine the transition probabilities $p(m, n \to i, j)$ for our Markov chain. One mode is the foreground mode during which the sofar unserviced (foreground) programs are allocated a quantum of service. Assuming the same population of programs as for the RR model, we then have the interval $(t_{k+1} - t_k)$ distributed as before according to the following distribution function

$$F(x) = Pr[(t_{k+1} - t_k) < x] = \begin{cases} 0; & x < \tau \\ 1 - \epsilon^{-\mu(x-\tau)}; & \tau \leq x < q + \tau. \\ 1; & x \geq q + \tau \end{cases} \tag{27}$$

Formally the foreground mode exists for all those states $(m, n)$ such that $m > 0$.

A second mode is the background mode. As pointed out earlier, service could be quantum-controlled in the background as well as the foreground. In the present model, however, we shall assume that *programs in the background* are run to completion and are not interrupted by arrivals subsequent to the time they commence operation. Clearly, such a choice assures a minimum in potential swap time (one swap for each background program). This also motivates our assumption that programs arriving to an empty system are switched immediately to the background and run to completion. The methods we use are easily modified to handle other choices for the background service discipline. Clearly, the background "inter-epoch" interval $(t_{k+1} - t_k)$ will be distributed as the sum of the constant swap time $\tau$ and an exponentially distributed random variable with parameter $\mu$. (This last statement is based, of course, on the memoryless property of the exponential distribution.) The background mode exists for states $(m, n)$ such that $m = 0$ and $n > 0$. The third mode of operation is obviously the idle mode for which (0, 0) is the only state, by definition. Using reasoning and notation similar to that for the RR model we may write the transition probabilities as follows:

Foreground mode: $m > 0$, $n \geq 0$,

$$p(m, n \to i, j) = \begin{cases} \displaystyle\int_0^q P(i - m + 1 | t + \tau) \mu \epsilon^{-\mu t} dt; & j = n, \ i \geq m - 1 \qquad (28) \\ \epsilon^{-\mu q} P(i - m + 1 | q + \tau); & j = n + 1, \ i \geq m - 1; \qquad (29) \end{cases}$$

Background mode: $m = 0$, $n > 0$,

$$p(0, n \to i, j) = \int_0^\infty P(i | t + \tau) \mu \epsilon^{-\mu t} dt; \quad j = n - 1, \ i \geq 0; \tag{30}$$

Idle mode: $m = n = 0$,

$$p(0, 0 \to i, 0) = \int_0^\infty P(i | t + \tau) \mu \epsilon^{-\mu t} dt; \quad i \geq 0, \tag{31}$$

where we have again used the memoryless property of the exponential distribution. We have $p(m, n \to i, j) = 0$ for all combinations of $(m, n)$ and $(i, j)$ not taken into account by Eqs. (28–31).

Assuming that the average input rate of programs is less than the maximum rate at which the computer can process them, an equilibrium probability distribution $\pi_{ij}$ for the states $(i,j)$ $(i=0,1,2,\ldots;$ $j=0,1,2,\ldots)$ will exist. By the definition of the equilibrium probability distribution we have

$$\pi_{ij}=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}p(m,n\rightarrow i,j)\pi_{mn}. \tag{32}$$

For the probabilities $\pi_{ij}$, we now define the following bivariate generating function

$$T(z_1,z_2)=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\pi_{ij}z_1^iz_2^j. \tag{33}$$

We may use Eq. (32) to obtain, after interchanging summations as before,

$$T(z_1,z_2)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\pi_{mn}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}p(m,n\rightarrow i,j)z_1^iz_2^j. \tag{34}$$

Separating Eq. (34) according to the different modes of operation and using the constraints on $i$ and $j$ gives

$$T(z_1,z_2)=\sum_{m=1}^{\infty}\sum_{n=0}^{\infty}\pi_{mn}\sum_{i=m-1}^{\infty}\sum_{j=n}^{n+1}p(m,n\rightarrow i,j)z_1^iz_2^j$$

$$+\sum_{n=1}^{\infty}\pi_{0n}z_2^{n-1}\sum_{i=0}^{\infty}p(0,n\rightarrow i,n-1)z_1^i+\pi_{00}\sum_{i=0}^{\infty}p(0,0\rightarrow i,0)z_1^i. \tag{35}$$

By retaining the functions $P(z)$ and $Q(z)$ as defined by Eq. (7) in the previous section and defining

$$R(z)=\sum_{i=0}^{\infty}z^i\int_0^{\infty}P(i|t+\tau)\mu\epsilon^{-\mu t}dt=\frac{\epsilon^{-\lambda\tau(1-z)}}{1+\rho(1-z)}, \tag{36}$$

we may simplify Eq. (35) to

$$T(z_1,z_2)=[T(z_1,z_2)-T_0(z_2)]\left[\frac{P(z_1)}{z_1}+\frac{z_2Q(z_1)}{z_1}\right]+[T_0(z_2)-\pi_{00}]\frac{R(z_1)}{z_2}+\pi_{00}R(z_1) \tag{37}$$

or

$$T(z_1,z_2)=\frac{[z_1R(z_1)-z_2P(z_1)-z_2^2Q(z_1)]T_0(z_2)-z_1R(z_1)(1-z_2)\pi_{00}}{z_2[z_1-P(z_1)-z_2Q(z_1)]}, \tag{38}$$

where

$$T_0(z_2)\equiv T(0,z_2)=\sum_{j=0}^{\infty}\pi_{0j}z_2^j \tag{39}$$

To complete our calculation of $T(z_1,z_2)$ we now set about finding expressions for $T_0(z_2)$ and $\pi_{00}$ appearing in Eq. (38). To find $\pi_{00}$, the probability of an empty system, from Eq. (38) we first find expressions for the foreground and background generating functions defined as follows:

$$T_f(z_1)=\lim_{z_2\rightarrow 1}T(z_1,z_2) \tag{40}$$

$$T_b(z_2)=\lim_{z_1\rightarrow 1}T(z_1,z_2). \tag{41}$$

By setting $z_2 = 1$ in Eq. (38), we have

$$T_f(z_1) = \frac{z_1 R(z_1) - [P(z_1) + Q(z_1)]}{z_1 - [P(z_1) + Q(z_1)]} \pi_{f0}, \tag{42}$$

where

$$\pi_{f0} = \lim_{z_2 \to 1} T_0(z_2) \tag{43}$$

is the probability that the system is *not* in the foreground mode of operation. To obtain an expression for $\pi_{f0}$ we now use

$$\lim_{z_1 \to 1} T_f(z_1) = 1.$$

Since $R(1) = P(1) + Q(1) = 1$ both numerator and denominator of Eq. (42) vanish at $z_1 = 1$. An application of l'Hospital's rule gives the following expression for $\pi_{f0}$:

$$\pi_{f0} = \frac{1 - [P'(1) + Q'(1)]}{1 + R'(1) - [P'(1) + Q'(1)]}. \tag{44}$$

From Eq. (36), we find

$$R'(1) = \rho(1 + \mu\tau), \tag{45}$$

so that in conjunction with Eqs. (12) and (14) of the previous section we have

$$\pi_{f0} = 1 - \frac{\rho(1 + \mu\tau)}{1 + \rho\delta}. \tag{46}$$

Now for $T_b(z_2)$, we set $z_1 = 1$ in Eq. (38) and obtain after simplification

$$T_b(z_2) = \frac{(1 + z_2\delta) T_0(z_2) - \pi_{00}}{z_2\delta}. \tag{47}$$

Using $\lim_{z_2 \to 1} T_b(z_2) = 1$, we obtain

$$\pi_{00} = (1 + \delta)\pi_{f0} - \delta. \tag{48}$$

Now substitution of Eq. (46) into Eq. (48) yields for $\pi_{00}$

$$\pi_{00} = 1 - \frac{\rho(1 + \delta)(1 + \mu\tau)}{1 + \rho\delta}. \tag{49}$$

In order to complete our calculation of $T(z_1, z_2)$, we must now find an expression for $T_0(z_2)$ as defined by Eq. (39). In the appendix we formally derive an expression for $T_0(z_2)$. A simple closed form solution has not been found, but the performance measures we shall obtain can be found without working directly with our expression for $T_0(z_2)$.

We now proceed to calculate the mean queue length $\bar{n}_f$ and $\bar{n}_b$ for the foreground and background, respectively. For $\bar{n}_f$ we proceed in the usual way by using

$$\lim_{z_1 \to 1} T_f'(z_1) = \bar{n}_f.$$

From Eq. (42) for $T_f(z)$ differentiation and taking the limit yield

$$\bar{n}_f = \frac{2(\rho + \lambda\tau)[1 - \lambda\tau - \rho\lambda q\delta - \rho(1 - \delta)(1 - \rho - \lambda\tau)] + (\lambda\rho)^2(1 + \rho\delta)}{[1 - \rho(1 - \delta) - \lambda\tau]^2} \pi_{f0}, \tag{50}$$

where $\pi_{f0}$ is given by Eq. (46) and $\delta = \epsilon^{-\mu q}$.

The evaluation of $\bar{n}_b$ is somewhat less straightforward since it involves

$$\lim_{z_2 \to 1} T_0(z_2)$$

which we cannot evaluate directly. We proceed as follows. First we use

$$\lim_{z_2 \to 1} T_b'(z_2) = \bar{n}_b$$

to obtain from Eq. (47)

$$\delta \bar{n}_b = (1+\delta)\bar{n}_b^0 - \pi_{f0} + \pi_{00}, \tag{51}$$

where

$$\bar{n}_b^0 = \lim_{z_2 \to 1} T_0'(z_2).$$

Now $\pi_{f0}$ and $\pi_{00}$ are known from Eqs. (46) and (49), so to obtain $\bar{n}_b$ explicitly, we must find $\bar{n}_b^0$. For this we shall find another expression from Eq. (38) that relates $\bar{n}_b$ and $\bar{n}_b^0$. First we rewrite Eq. (38) as

$$A(z_1, z_2)T(z_1, z_2) = B(z_1, z_2)T_0(z_2) + C(z_1, z_2)\pi_{00}, \tag{52}$$

where

$$\begin{aligned} A(z_1, z_2) &= z_2[z_1 - P(z_1) - z_2 Q(z_1)] \\ B(z_1, z_2) &= z_1 R(z_1) - z_2 P(z_1) - z_2^2 Q(z_1) \\ C(z_1, z_2) &= z_1(z_2 - 1)R(z_1). \end{aligned} \tag{53}$$

We now calculate the mixed derivative

$$\frac{\partial^2 T(z_1, z_2)}{\partial z_1 \partial z_2}.$$

Since $A(1, 1) = 0$ we note that the term with the mixed derivative will vanish when we let $z_1$ and $z_2$ approach one in the resulting expression. This leaves us, as shown below, an expression relating $\bar{n}_b$, $\bar{n}_b^0$, and known quantities. This will be the expression we are after. Carrying out the differentiations gives

$$\begin{aligned} &\bar{n}_b \frac{\partial A}{\partial z_1}\bigg|_{z_1=z_2=1} + \bar{n}_f \frac{\partial A}{\partial z_2}\bigg|_{z_1=z_2=1} + \frac{\partial^2 A}{\partial z_1 \partial z_2}\bigg|_{z_1=z_2=1} \\ &\qquad = \pi_{f0} \frac{\partial^2 B}{\partial z_1 \partial z_2}\bigg|_{z_1=z_2=1} + \bar{n}_b^0 \frac{\partial B}{\partial z_1}\bigg|_{z_1=z_2=1} + \pi_{00} \frac{\partial^2 C}{\partial z_1 \partial z_2}\bigg|_{z_1=z_2=1}. \end{aligned} \tag{54}$$

Evaluation of the partial derivatives yields the equation

$$\begin{aligned} &[1-\rho(1-\delta)-\lambda\tau]\bar{n}_b - (1+\rho\delta)\bar{n}_b^0 + [1-\rho(1-\delta)-\lambda\tau(1+\delta)-\lambda q\delta] \\ &\qquad = \delta\bar{n}_f - [\rho(1-\delta)+\lambda\tau(1+\delta)+\lambda q\delta]\pi_{f0} + (1+\rho+\lambda\tau)\pi_{00}. \end{aligned} \tag{55}$$

By using Eq. (55) to eliminate $\bar{n}_b^0$ from Eq. (51), we finally have

$$\bar{n}_b = \frac{\delta(1+\delta)\bar{n}_f + [1+\rho\delta-(1+\delta)(1+\rho+\lambda\tau)](\pi_{00}-\pi_{f0}) - (1+\delta)[1+\rho\delta-\lambda\delta(q+r)](1-\pi_{f0}) + \rho(1+\mu\tau)(1+\delta)}{1-[\rho-\lambda\tau(1+\delta)]}, \tag{56}$$

where $\pi_{f0}$, $\pi_{00}$, and $\bar{n}_f$ are given by Eqs. (46), (49), and (50), respectively.

From Eq. (50), we see that the condition for the equilibrium solution in *foreground* operations is

$$\lambda[(1/\mu)(1-\delta)+\tau]<1. \tag{57}$$

Since the bracketed quantity is simply the mean of the foreground running times (i.e. the mean of $F(x)$ in Eq. (27)) the above condition reduces to $\rho_f<1$, where $\rho_f=\rho(1-\delta)+\lambda\tau$ is the foreground utilization factor. Now in Eq. (56) there are two poles. Because of the appearance of $\bar{n}_f$ in Eq. (50) the first one corresponds to the one for $\bar{n}_f$ given above. The second corresponds to the zero of the denominator when $\rho+\lambda\tau(1+\delta)=1$. Thus we have the additional condition for equilibrium in the background

$$\lambda[1/\mu+\tau(1+\delta)]<1. \tag{58}$$

To interpret the bracketed expression, we see that it is the mean $(1/\mu)$ of a program service time, plus the swap time $\tau$ experienced in the foreground, plus the swap time experienced in the background ($\tau$ again) times the probability $\delta=\epsilon^{-\mu q}$ that a program requires background operation. Now for all $q>0$, $0<\delta=\epsilon^{-\mu q}<1$, and with this constraint on $\delta$ it is easy to show that $\rho_f<\rho+\lambda\tau(1+\delta)$ for all $q>0$. (See Figure 3.) Thus the quantity $\rho_S=\rho+\lambda\tau(1+\delta)$ may be viewed as the utilization factor for

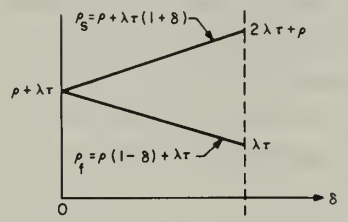



FIGURE 3. Comparison of $\rho_f$ and $\rho_s$.

the *system*, and for equilibrium of the system $\rho_S<1$. That is, the equilibrium of the system as a whole is governed by the background, which becomes saturated before the foreground. Qualitatively, the effect is that when the foreground starts to become too busy (but still short of foreground saturation) the background does not have sufficient time to operate its programs and becomes saturated. The above remarks will be illustrated later on in a section containing some examples.

**Conditional Waiting Times**

We proceed to find in this section a mean (conditional) waiting time result similar to that derived for the RR model earlier. That is, we consider the mean elapsed time between the regeneration point following the arrival time of a program requiring $s$ secs of service and the instant that the program reaches the service point for the last time. Again let $W_s$ be this mean waiting time. We have for $W_s$ the following expression

$$W_s=\begin{cases}\bar{n}_f m_s;\ s\leq q\\ \bar{n}_f m_s+\dfrac{(\bar{n}_f m_s+q+\tau)\lambda m_s+(\delta\bar{n}_f+\bar{n}_b)(1/\mu)}{1-\lambda m_s}.\end{cases} \tag{59}$$

For the proof, we divide into four intervals $\tau_i$, $i=1,2,3,4$, the waiting time of an arriving program (to be called the *tagged* program) whose processing requirements exceed $q$ secs. The first interval is simply the time spent waiting in the foreground queue for its first quantum of service. Clearly, the mean $\bar{\tau}$ of this interval will be the average waiting time in queue of a program requiring less than or equal to $q$ secs of service. Suppose the arriving (tagged) program encounters $n_f$ programs in the foreground and $n_b$ programs in the background. (The program, if any, that is just commencing service is

included in $n_f$ or $n_b$, depending on the mode of operation.) Then the tagged program must wait through $n_f$ foreground operations until it reaches the service point. Therefore, due to independence among service times, taking expected values

$$\tau = \bar{n}_f m_s. \tag{60}$$

The second interval $\tau_2$ is the busy period of foreground operations following the interval $\tau_1 + q + \tau$. The time $q+\tau$ is added to $\tau_1$ because of the tagged program's operation. From the definition of $\lambda$ we note that the average number of programs with which this busy period starts is $\lambda(\tau_1 + q + \tau)$. To determine the average length of this busy period we use the following result which Takacs [18] states and proves as a theorem. Let $\zeta$ denote the mean busy period (commencing with one unit) for a single channel queueing system with Poisson input and general service time. Then,

$$\zeta = \frac{a}{1 - \lambda a}, \tag{61}$$

where $\lambda$ is the intensity (average arrival rate) of the Poisson input and $a$ is the mean service time. Now we want the average busy period beginning with $\lambda(\tau_1 + q + \tau)$ programs rather than just one program. But this will simply be $\lambda(\tau_1 + q + \tau)\zeta$ where $\zeta$ is given by Eq. (61). To justify this assertion we reason as follows. During the interval $\tau_2$ we suppose the following queue discipline in effect for the foreground programs. After the interval $\tau_1 + q + \tau$ we allow the first arrival to operate. Before we operate the second program that arrived during the interval $\tau_1 + q + \tau$ we process all new arrivals until none remain. We then allow the second arrival during $\tau_1 + q + \tau$ to be serviced and process subsequent arrivals as before We continue in this way until all of the arrivals during $\tau_1 + q + \tau$ and all subsequent arrivals as described above, have been serviced. Since the busy period in question is unaffected by the sequence in which foreground programs are serviced we see that a busy period commencing with $k$ programs is equivalent to $k$ disjoint and independent busy periods commencing with one program. Finally, therefore, substituting $m_s$ for $a$ in Eq. (61) and using this last result gives

$$\bar{\tau}_2 = \frac{\lambda m_s}{1 - \lambda m_s} (\bar{\tau}_1 + q + \tau). \tag{62}$$

The third interval $\tau_3$ is made up of the complete operation times of those background programs ahead of the tagged program when the latter reaches the background queue. The number of these programs will be given by the number of programs passing from the $n_f$ foreground programs to the background plus the number originally in the background queue. Thus, the mean number of background programs ahead of the tagged program will be $\delta\bar{n}_f + \bar{n}_b$, where $\delta$ is the probability that a foreground program passes into the background. From the memoryless property of the exponential distribution we know that the mean service time of a program in the background is $1/\mu$. Thus

$$\bar{\tau}_3 = (\delta\bar{n}_f + \bar{n}_b)(1/\mu) \tag{63}$$

Finally, the fourth interval $\tau_4$ consists of the foreground busy periods that occur during the background operations ($\tau_3$). These busy periods are due, of course, to new arrivals. Although these busy periods actually alternate with background operations, for the purposes of computing $\bar{\tau}_4$ we may consider them as combined into one large busy period beginning with an average of $\lambda\bar{\tau}_3$ programs. From Eq. (61), therefore, we obtain

$$\bar{\tau}_4 = \frac{m_s\lambda}{1-m_s\lambda}\,\bar{\tau}_3. \tag{64}$$

Adding Eqs. (60), (62), (63), and (64) for the $\bar{\tau}_i$, we obtain after simplification

$$W_s = \bar{n}_f m_s + \frac{(\bar{n}_f m_s + q + \tau)\lambda m_s + (\delta\bar{n}_f + \bar{n}_b)(1/\mu)}{1-\lambda m_s}.$$

## III. EXAMPLES AND DISCUSSION

This section commences with a brief treatment of the generalization to random swap times. Subsequently, a number of examples of system behavior are discussed. In particular, mean waiting times and mean numbers in the queue are investigated as they vary with changes in the system parameters.

The power of the imbedded-Markov-chain approach is exhibited by the amount of structural detail of the system being modeled that can be included in the analysis. Consider the following useful example. Let us suppose that in the RR or FB model we wish to analyze overhead effects in more detail. In particular, assume that the overhead parameter $\tau$ is to be specified as $\tau=\tau_0+\tau_s$ where, in the process of switching programs, $\tau_0$ is the operating time (assumed constant) of the control or executive program mentioned earlier and $\tau_s$ is the swap time. Furthermore, assume that $\tau_s$ is a random variable subject to some arbitrary distribution function $S(\tau_s)$ depending on the input/output devices involved. $S(\tau_s)$ may also include some consideration of possible overlapping in swapping and program execution. Focusing on the RR model as a specific illustration we see that the transition probabilities of Eq. (3) must be integrated over the distribution of $\tau_s$. Thus, identifying our new transition probabilities by an asterisk, we have

$$p_{ij}^* = \int_0^\infty p_{ij}(\tau_0+\tau_s)\,dS(\tau_s).$$

The analysis is then carried out for the $p_{ij}^*$ as shown in Section II.

Our first example in Figure 4 illustrates the effect of $\tau$, the swap time, on the mean congestion (mean number in the system) in the round-robin system. Equation (18) of Section II is plotted in Figure 4. The parameter values chosen are $\lambda=1.0$/sec, $\mu=2.5$ sec, and three values for the quantum size:

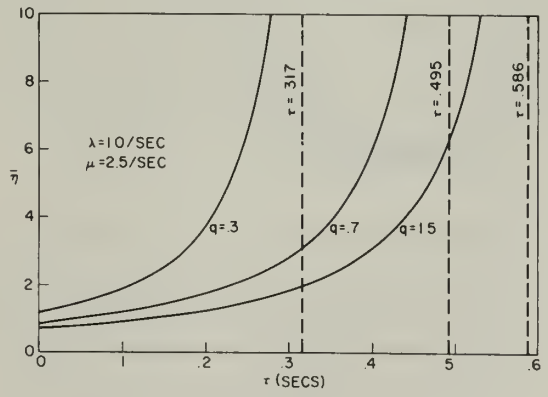



FIGURE 4. Round-Robin Mean *vs* Swap Time.

$q=0.3$, 0.7, 1.5 secs. It can be verified from the expression for $\beta''(1)$ in Eq. (17) that it is analytic at $\pi_0=0$. Thus, from Eq. (18) the constraint on the loading for the RR system that must be satisfied for equilibrium operation can be written as

$$(1-\rho)(1-\delta)-\lambda\tau>0.$$

This expression may be rendered more conveniently as

$$\rho_{RR} \equiv \lambda\left[(1/\mu)+\frac{\tau}{1-\delta}\right] < 1, \tag{65}$$

where $\rho_{RR}$ may be interpreted as the utilization factor for the RR system. To interpret the bracketed quantity, we first observe that the probability of a program requiring exactly $k$ quanta of service is $(1-\delta)\delta^{k-1}$ (the geometric distribution). From this we see that

$$\frac{1}{1-\delta} = (1-\delta)\sum_{k=1}^{\infty} k\delta^{k-1}$$

is the mean number of quanta required by the programs arriving to the RR system, and that $\tau/1-\delta$ is just the mean amount of swap time required by a program. Finally, therefore, $(1/\mu)+\tau/1-\delta$ simply represents the expected amount of time required by a program in the RR system, which brings $\rho_{RR}$ in accord with the usual definition of the utilization factor for a queueing system (i.e., the average input rate divided by the maximum output rate). From both Figure 4 and Eq. (65) it can be seen that an increase in the quantum increases the maximum loading that the RR system is able to sustain. The asymptotes for the curves in Figure 4 are easily calculated by setting $\rho_{RR}=1$, substituting the parameter values, and solving for $\tau$. The values of the asymptotes along which $\bar{n}$ approaches infinity are shown in the illustration.

In Figure 5, the foreground and background means ($\bar{n}_f$ and $\bar{n}_b$) for the FB system are plotted versus the loading ($\rho$) for several values of the swap time $\tau$ (the first-come-first-served curve is also

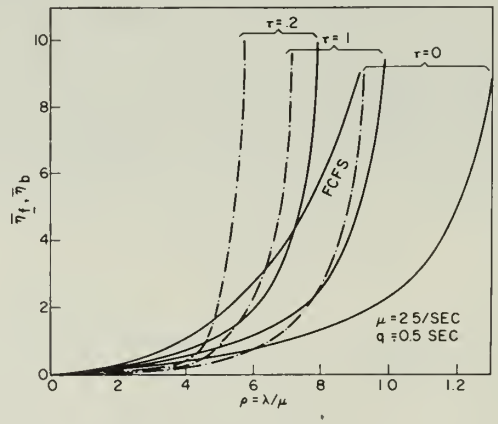



FIGURE 5. The Foreground Mean ($\bar{n}_f$) and Background Mean ($\bar{n}_b$) *vs* $\rho=\lambda/\mu$.

shown). The parameters chosen are $q=0.5$ sec, $\mu=2.5$ sec, and $\tau=0$, 0.1, and 0.2 sec. (The curves for $\bar{n}_b$ are given in dashed lines.) The asymptotes for the curves shown are found from the foreground and system (background) utilization factors discussed in Section II. Figure 5 bears out the fact, discussed in Section II, that the constraint $\rho_S < 1$ governs when the system as a whole becomes saturated; however, between $\rho_f=1$ and $\rho_S=1$ exists a region that corresponds to a system providing an infinite expected wait for background programs and a finite expected wait for programs requiring less than one quantum of service. Note, from Figure 5, that background congestion is less than that of the foreground only for small loading.

In Figure 5, we have chosen a quantum (0.5 sec) somewhat larger than the mean operation time ($1/\mu=0.4$ sec). Figure 6 shows what happens when we select a quantum size (0.1 sec) substantially

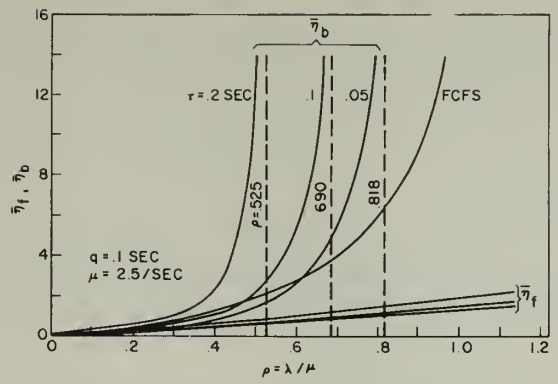



FIGURE 6. The Foreground Mean ($\bar{n}_f$) and Background Mean ($\bar{n}_b$) *vs* $\rho = \lambda/\mu$.

less than the mean operation time. Again we have shown the first-come-first-served curve for reference. As can be seen, the region between foreground and background saturation is very much larger; however, the average number of programs whose service requirement is met by one quantum in the foreground is far less than for the quantum (0.5 sec) of Figure 5. For reference, we have also shown in Figure 6 the first-come-first-served mean given by $\rho/(1-\rho)$. The asymptotes along which the curves of Figures 5 and 6 approach infinity are calculated from Eqs. (57) and (58) after inserting the parameter values for $q$, $\mu$, and $\tau$.

A common measure of the waiting time response of a time-sharing system with an RR discipline is the so-called "cycle" time [7], [8], [15] Although there are several possible definitions we shall define the mean cycle time, $W_c$, as the average time to process for one quantum the mean number in the system ($\bar{n}$) in equilibrium. Thus, setting $m=1$ in Eq. (21) of Section II, we obtain

$$W_c = \bar{n} m_s, \tag{66}$$

where $m_s$ is the mean operation time of a program, within the quantum constraint. We have

$$m_s = (1/\mu)(1-\delta) + \tau. \tag{67}$$

Figure 7 plots $W_c$ versus the quantum size $q$ for various values of $\tau$, with constant loading: $\lambda = 1.0$/sec, and $\mu = 2.5$/sec. It is evident that a quantum less than 0.5 sec for $\tau = 0.2$ sec, 0.3 sec for $\tau = 0.15$ sec, and 0.15 sec for $\tau = 0.1$ sec causes the waiting time performance to deteriorate sharply. As $q$ becomes

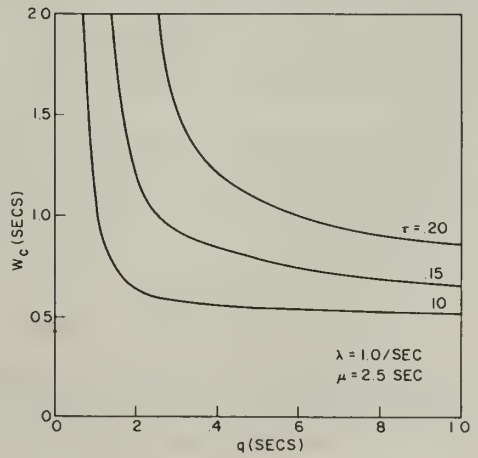



FIGURE 7. Cycle Time *vs* Quantum Size.

small a program must be swapped a large number of times in order to receive its total service requirement. The increase in swapping overhead thus increases the mean number in the system. From Eq. (67) the mean operation time becomes dominated by the swap time $\tau$ for small $q$, so that $W_c$ increases with the increase in $\bar{n}$. On the other hand, it is useful to note that the curves become quite flat beyond these points (especially for $\tau < 0.15$), so that over-specification of quantum size is not as seriously detrimental as is under-specification.

For our last example, we investigate waiting times for foreground and background programs in the FB system as they are made to vary with changes in the quantum size and swap time. The waiting times of foreground programs requiring but one quantum of service are plotted in Figure 8 versus the quantum

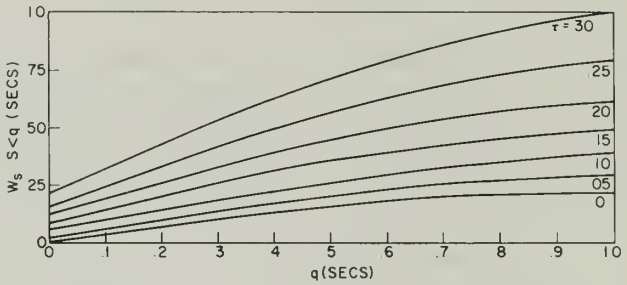



FIGURE 8. Foreground Waiting Times *vs* Quantum Size.

size, $q$, for values of $\tau$ ranging from $\tau = 0.5$ to $\tau = 3.0$. The remaining parameters were chosen as $\lambda = 1.0$/sec and $\mu = 2.5$/sec. It should be noted that as the quantum size increases, the class of foreground programs (those requiring less than $q$ secs of service) also enlarges; thus the waiting time for any given foreground program increases. If we fix on any given foreground program size then, taking 0.5 sec as an example, we may look at the curves of Figure 8 only to the right of the 0.5 sec ordinate to see what happens as $q$ increases. Clearly, if the quantum size is less than 0.5 secs, the given program will no longer be a foreground program. In the limit as $q$ goes to infinity, the curves of Figure 8 approach a value which corresponds to a first-come-first-served system in which there exists a fixed, initial loading and final unloading cost lumped into the parameter $\tau$. In the limit, of course all programs become foreground programs.

In Figure 9 are shown the mean waiting times for background programs for the same values of the parameters $\lambda$, $\mu$, and $\tau$. Here again, when considering a specific program running time we must look at only part of each of the several curves; this time, however, we can consider only those quantum sizes to the left of the given running time. Those values to the right would make the program a foreground program whose mean waiting time would be described by Figure 8. Of particular interest here are the optimum points that exist. For $\tau < 0.2$ the optimum point for a *given* program is that value of $q$ corresponding to the program's running time. For any $t \geq 0.2$, however, this is not true for programs whose running times lay between the low point of the corresponding curve and $q = 0$. In this region too many programs pass into the background (because of the small quantum) and incur the large swap time in so doing; that is, the value of Eq. (59) in Section II for the background waiting time is dominated by the large swap time. Therefore, an increase in quantum size reduces waiting times by reducing the number of swaps necessary. To the right of all the optimum points the mean waiting time must increase with $q$ since more and more programs must be processed which arrive during a background program's first quantum of operation. It must be emphasized, however, that the

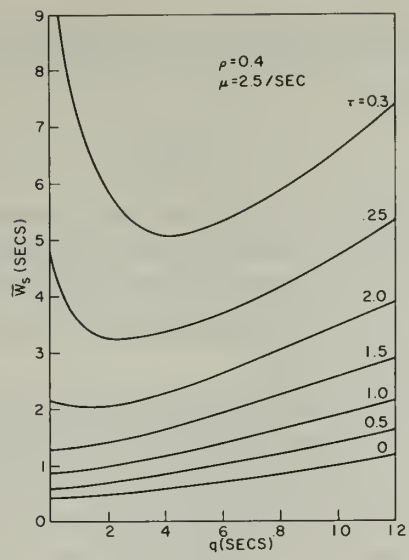



FIGURE 9. Background Waiting Times *vs* Quantum Size.

farther to the right one proceeds the smaller the class of programs to which the curves are applicable. That is, the curves are applicable only to a class of larger and larger (running time) programs.

**Conclusions**

Again, the basic objective has been to provide an analysis by which the effects of overhead costs (lumped into the swap time parameter) and quantum size can be assessed with respect to different quantum-controlled service disciplines. Specifically, the effects have been observed on the waiting time and congestion (mean number in the system) performance measures, although the methods presented can be employed for studying the effects on other performance measures (e.g., the variance of the queue-length distribution).

The extensions to the analysis presented in this paper, constituting suggestions for further study, coincide precisely with the constraints and limiting assumptions that have been made. Accordingly, we may list the more important suggestions for further study as follows:

1. Removal of the limitation of the results of the Markov chain analysis to discrete time instants (see Ref. [7] for an example involving renewal theory in which a similar generalization has been made).
2. Removal of the restriction to the exponential distribution for servicing times in both the round-robin and foreground-background models.
3. Removal of the limitation to two levels in the foreground-background model.
4. Development of more usable expressions for the waiting time distributions for the round-robin and foreground-background models.

## Appendix

### CALCULATION OF $T_0(z_2)$ AND FINAL FORM OF $T(z_1, z_2)$

To find $T_0(z_2)$, we first note that

$$\pi_{0n} = \pi_n^0 \, \pi_{f0},$$

where $\pi_n^0$ is the probability there are $n$ programs in the background conditioned on there being none in the foreground. Thus

$$T_0(z_2) = \pi_{f0} \sum_{n=0}^{\infty} \pi_n^0 z_2^n. \tag{A1}$$

Now we find the generating function for the $\pi_n^0$ by analyzing a second chain imbedded in the former Markov chain $\xi(t_k)$. The second chain is defined by observing the state of the system only at those epochs of $\xi(t_k)$ at which there are no foreground programs in the system. The state of our new chain will simply be the number of programs in the background. For reasons identical to those given earlier for $\xi(t_k)$ the new chain will be a Markov chain.

We now proceed to find the transition probabilities $q_{ij}$; that is, the probability of passing from $i$ programs (in the background) to $j$ programs (in the background) in one transition. In computing $q_{ij}$ we observe that in the transition $i \to j$ a busy period may or may not intervene depending on whether or not any arrivals occurred during the time of the transition. Here, we define a busy period as a period of operation during which the system is in the foreground mode. For the case $j \geq i > 0$, for which there must be an intervening busy period, we note first that the busy period may commence with from one to an infinite number of programs depending on the number of arrivals during the preceding background operation. Let $f_k(n)$ be the probability there are $n$ services in a busy period that begins with $k$ programs. (Clearly, $f_k(n) = 0$ for $n < k$.) Now out of $n$ programs the probability that $m$ of them require in excess of $q$ secs of service (and therefore will be background programs) is given by the binomial distribution;

$$\binom{n}{m} \delta^m (1-\delta)^{n-m}.$$

For the transition $i \to j$, we require $j - i + 1$ background arrivals if $j \geq i$. Thus we may write

$$Pr[i \to j | k] = \sum_{n=j-i+1}^{\infty} \binom{n}{j-i+1} \delta^{j-i+1} (1-\delta)^{n-(j-i+1)} f_k(n) \ , \tag{A2}$$

where Pr $[i \to j|k]$ is the probability of the transition $i \to j$ given that $k$ arrivals occurred during the operation of a background program. Let $u(k)$ denote the distribution for $k$. Then we have

$$u(k) = \int_0^{\infty} P(k|t+\tau) \mu \epsilon^{-\mu t} dt = \epsilon^{-\lambda \tau} \int_0^{\infty} \frac{[\lambda(t+\tau)]^k}{k!} \mu \epsilon^{-(\mu+\lambda)t} dt. \tag{A3}$$

Finally, therefore, we may write

$$q_{ij} = \sum_{k=1}^{\infty} u(k) Pr[i \to j|k],$$

whereupon substitution of Eqs. (A2) and (A3) yields,

$$q_{ij} = \sum_{k=1}^{\infty} \epsilon^{-\lambda \tau} \int_0^{\infty} \frac{[\lambda(t+\tau)]^k}{k!} \mu \epsilon^{-(\mu+\lambda)t} dt \times \sum_{n=j-i+1}^{\infty} \binom{n}{j-i+1} \delta^{j-i+1} (1-\delta)^{n-(j-i+1)} f_k(n); \ i > 0, j \geq i. \tag{A4}$$

Now $f_k(n)$ may be determined as follows. First of all Takacs [18] has shown that if $k = 1$, and if

$$A(s) = \sum_{i=0}^{\infty} f_1(i) s^i \tag{A5}$$

denotes the generating function for the $f_1(i)$, then the $f_1(i)$ are determined by the following integral equation

$$A(s) = s \int_0^{\infty} \epsilon^{-\lambda t[1 - A(t)]} dF(t), \tag{A6}$$

where $F(t)$ is given by Eq. (27). It is easy to show (see Takacs [18]) that the distribution of a busy period beginning with $k$ programs is simply the $k$-fold convolution of the distribution corresponding to a busy period beginning with one program. Similarly, then, we see that $f_n(n)$ is the $k$-fold convolution (in the discrete sense) of $f_1(n)$ with itself.

For $j=i-1\geqslant 0$ we may or may not have an intervening busy period. Given an intervening busy period the probability of the transition $i\to i-1$ is given by Eq. (A4) above with $j=i-1$. For no intervening busy period (i.e., no arrivals), the above transition probability is given by the expression;

$$\int_0^\infty P(0|t+\tau)\mu\epsilon^{-\mu t}\,dt=\frac{\epsilon^{-\lambda\tau}}{1+\rho} \tag{A7}$$

Thus, $q_{i,i-1}$ will be given by the sum of Eq. (A7) and Eq. (A4) with $j=i-1$.

Our calculation of $q_{ij}$ is completed by observing that the transition $0\to j$ involves considerations only of busy periods beginning with one program (the first arrival). Thus

$$q_{0j}=\sum_{n=1}^{\infty} f_1(n)\binom{n}{j}\delta^j(1-\delta)^{n-j}. \tag{A8}$$

For the equilibrium probability distribution (of the second chain), we have as before

$$\pi_n^0=\sum_{m=0}^{\infty} q_{mn}\pi_m^0, \tag{A9}$$

so that forming the generating function gives for Eq. (A1)

$$T_0(z_2)=\pi_{f0}\sum_{m=0}^{\infty}\pi_m^0\sum_{n=0}^{\infty} q_{mn}z_2^n. \tag{A10}$$

Now let

$$P_1(z_2)=\sum_{k=0}^{\infty} q_{i,i+k}z_2^k, \tag{A11}$$

$$Q_1(z_2)=\sum_{k=0}^{\infty} q_{0k}z_2^k, \tag{A12}$$

and

$$R_1=q_{i,i-1}. \tag{A13}$$

Substitution of Eqs. (A11–A13) into Eq. (A10) yields after simplification

$$T_0(z_2)=\frac{Q_1(z_2)-P_1(z_2)-R_1z_2^{-1}}{1-P(z_2)-R_1z_2^{-1}}\,\pi_{00}. \tag{A14}$$

Substituting Eq. (A14) into Eq. (38) finally gives the general expression for $T(z_1, z_2)$ that we have been seeking

$$\begin{aligned}T(z_1, z_2)=\pi_{00}\{&[R(z_1)z_1z_2^{-1}-P(z_1)-z_2Q(z_1)][Q_1(z_2)-P_1(z_2)-R_1z_2^{-1}]\\&-[1-R_1z_2^{-1}-P_1(z_2)][z_1z_2^{-1}R(z_1)(1-z_2)]\}\\&\times\{[z_1-P(z_1)-z_2Q(z_1)][1-P_1(z_2)-R_1z_2^{-1}]\}^{-1}.\end{aligned} \tag{A15}$$

## REFERENCES

[1] Coffman, E. G., *Stochastic Models of Multiple and Time-Shared Computer Operation*, Ph.D. Dissertation, Dept. of Eng'g., University of California, Los Angeles, California, August 1966.

[2] Coffman, E. G. and B. Krishnamoorthi, "Preliminary Analyses of Time-Shared Computer Operation," *System Development Corp.*, SP–1719, August 1964.

[3] Corbato, F. T., M. Merwyn-Daggett and R. C. Daley, "An Experimental Time-Sharing System," *Spring Joint Computer Conference Proc.*, The National Press, Palo Alto, Calif., pp. 335–344, May 1962.

[4] Kendall, D. G., "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain," *Ann. Math. Stat.*, Vol. 24, pp. 338–354, September 1953.

[5] Kleinrock, Leonard, "Analysis of a Time-Shared Processor," *Naval Res. Logistics Quart.*, Vol. 11, pp. 59–73, March 1964.

[6] Kleinrock, Leonard, "Time-Shared Systems: A Theoretical Treatment," *Journal of the ACM*, Vol. 14, pp. 242–261, April 1967.

[7] Krishnamoorthi, B., "The Stationary Behavior of a Time-Sharing System Under Poisson Assumptions," *System Development Corp.*, SP–2090, September 1965.

[8] Krishnamoorthi, B. and R. C. Wood, "Time-Shared Computer Operations with Both Inter-arrival and Service Times Exponential," *Journal of the ACM*, Vol. 13, pp. 317–338, July 1966.

[9] McCarthy, J. S., E. Boilen, and J. C. R. Licklider, "A Time-Sharing Debugging System for a Small Computer," *Spring Joint Computer Conference Proc.*, Spartan Books, Baltimore, Md., pp. 51–57, May 1963.

[10] Miller, R. G., Jr., "Priority Queues," *Ann. Math. Stat.*, Vol. 31, pp. 86–193, March 1960.

[11] Patel, N. R., "A Mathematical Analysis of Computer Time-Sharing Systems," Interim Tech. Rep. No. 20, Army Res. Office (Durham), Grant No., DA–ARO(D)–31–124–G158 Oper. Res. Center, M.I.T., 1964.

[12] Phipps, J. E., Jr., "Machine Repair as a Priority Waiting-Line Problem," *Operations Research*, Vol. 4, pp. 76–85, 1956.

[13] Saaty, T. L., *Elements of Queueing Theory*, McGraw-Hill, New York, 1961.

[14] Scherr, Allen L., *An Analysis of Time-Shared Computer Systems*, Ph.D. Dissertation, Dept. of Elect. Eng'g., Massachusetts Institute of Technology, Cambridge, Mass., June 1965.

[15] Schrage, L. E., *Some Queueing Models for a Time-Shared Facility*, Ph.D. Dissertation, Dept. of Indust. Eng'g., Cornell University, Ithaca, N.Y., Feb. 1966.

[16] Schwartz, J. I., E. G. Coffman, and C. Weissman, "A General Purpose Time-Sharing System." *Spring Joint Computer Conference Proc.*, Spartan Books, Baltimore, Md., pp. 347–411, May 1964.

[17] Shemer, J. E., "Some Mathematical Considerations of Time-Sharing Scheduling Algorithms," *Journal of the ACM*, Vol. 14, pp. 262–272.

[18] Takacs, L., *Introduction to the Theory of Queues* (Oxford University Press, New York, 1962).

[19] Takacs, L., "Priority Queues," *Operations Research*, Vol. 12, pp. 63–74. Jan.–Feb. 1964.

## BIBLIOGRAPHY

Corbato, F. T., and V. A. Vyssotsky, "Introduction and Overview of the Multics Systems," *Proceedings Fall Joint Computer Conference*, Spartan Books, Washington, D.C., Vol. 27, pp. 185–196, 1965.

Keilson, J. and A. Kooharian, "On Time-Dependent Queueing Processes," *Ann. Math. Stat.*, Vol. 31, pp. 104–112, March 1960.

Keston, H. and J.Th. Runnenburg, "Priority in Waiting Line Problems," *Koninkl, Ned. Akad. Wetenschap. Proc.*, Ser. A., Vol. 60, pp. 312–324, (Part I) and pp. 325–336 (Part II), 1957.

* * *

# UNIFORMLY MINIMUM VARIANCE UNBIASED ESTIMATES OF OPERATIONAL READINESS AND RELIABILITY IN A TWO-STATE SYSTEM

M. Mazumdar

*Westinghouse Research Laboratories*
*Pittsburgh, Pennsylvania*

**ABSTRACT**

This paper obtains the uniformly minimum variance unbiased estimates of two indices of performance of a system which alternates between two states "up" or "down" in accordance with a Markov process. The two indices are (1) operational readiness, which measures the probability that the system will be up when needed; and (2) operational reliability, which measures the probability that the system will be up during the entire time of need. For the purpose of obtaining these estimates, two types of observations are considered: (a) those which reveal only the state of system at isolated time-points, and (b) those which continuously record the duration of the "up" and "down" times of the system.

## 1. INTRODUCTION

In [1], we recently considered the problem of estimation of some reliability measures in a two state system which alternates between two capability states, "up" (operative) or "down" (inoperative) in accordance with a two-state Markov process. We remarked that such systems frequently occur in problems of system reliability, work sampling, and communications. We obtained the maximum likelihood estimates of some indices of performance of these systems and studied their asymptotic and small-sample properties. In this paper, we obtain the uniformly minimum variance unbiased (UMVU) estimates of two measures of performance in such systems using some of the sampling procedures described in [1]. The reliability measures for which estimators are derived here are: (i) *operational readiness*, which measures the probability that the system will be up when needed; and (ii) *operational reliability*, which measures the probability that the system will be up during the entire time of need. Section 2 of this paper describes in brief the system model and the sampling procedure. Sections 3 and 4 give the results. For a detailed description of the estimation problem considered here, the reader is referred to [1].

## 2. SYSTEM MODEL AND THE SAMPLING PROCEDURE

### 2.1. System Model

We assume that the times spent by the system in the "up" and "down" states are mutually independent and identically distributed, each having the exponential distribution. More specifically, if $U_i$ and $D_i$ are, respectively, the $i$-th "up and down times," then their densities are given by

$$f_U(x) = \mu e^{-\mu x} \qquad (x \geq 0), \tag{1}$$

and

$$f_D(y) = \lambda e^{-\lambda y} \qquad (y \geq 0),$$

where $\lambda$ and $\mu$ are positive parameters. In terms of these parameters, the above two reliability measures are given by

**Operational Readiness:**

$$p = Pr\{\text{up in the long run}\} = \frac{\lambda}{\lambda+\mu}, \text{ and} \tag{2}$$

**Operational Reliability:**

$$\rho(t) = \frac{\lambda}{\lambda+\mu} e^{-\mu t}. \tag{3}$$

Here $\rho(t)$ measures the long-run probability that the system will be up, say at the start of a mission, and remain so for the mission time $t$.

In this paper, we shall mainly consider sampling procedures which yield a sequence of observations (made continuously on intervals) on the "up" and "down" times of the system. These observations are called patches. The *UMVU* estimates corresponding to this type of observation are obtained in Section 3 and some limited comparisons are made between them and the maximum likelihood estimates. In [1] it was pointed out that the efficiency of estimates improves considerably if in addition to patches, isolated observations are taken at widely dispersed instants which reveal the state of the system at the instant of observation. These observations were called "snapshots." In Section 4, *UMVU* estimates are derived for $p$ in a special case when both snapshots and patches are used in the sampling procedure.

We now define the quantities which arise when one follows the above sampling procedure (at this stage, we keep our formulation sufficiently general so as to include all sampling plans in which patches and snapshots are mutually interspersed):

a: the number of times the system is up at the start of a patch;

$a$ may be fixed or random depending on the sampling plan.

$X_+$: total uptime observed; i.e.,

$$X_+ = \sum_{i=1}^{a} X_i,$$

$X_i$ being an individual up interval (or in some cases, a random modification thereof).

b: the number of times the system is down at the start of a patch;

$b$ may be fixed or random depending on the sampling plan.

$Y_+$: total downtime observed, i.e.,

$$Y_+ = \sum_{i=1}^{b} Y_i,$$

$Y_i$ being an individual down interval (or in some cases, a random modification thereof).

$\alpha$: total number of snapshots showing the system to be up.

$\beta$: total number of snapshots showing the system to be down.

When the observations are made at widely spaced instants, in which case the states observed are effectively independent, the probability that the system is observed up is $p$. It follows from the above description that the likelihood function of the observations following the above sampling procedure will be given by

$$L(\lambda;\mu)=e^{-\mu X_+}\mu^a e^{-\lambda Y_+}\lambda^b\left(\frac{\lambda}{\lambda+\mu}\right)^\alpha\left(\frac{\mu}{\lambda+\mu}\right)^\beta. \tag{4}$$

While implementing the sampling procedure in practice, however, one will invariably be faced with the following question: "how far should the 'snapshot' observations be dispersed from each other so that they are effectively independent?" If $\lambda$ and $\mu$ were known, a sampling interval greater than, say $5(\lambda+\mu)^{-1}$ will ensure that (4) will approximately hold true; however, a statistician will often have at most a diffuse prior knowledge about the values of $\lambda$ and $\mu$. Thus there arises the following problem: "on the basis of a given prior distribution for $\lambda$ and $\mu$, how should one space the snapshots in a sequential manner?" In this paper, we do not consider this design problem and assume that the sampling interval is sufficiently large in comparison to the length of an individual "up" and "down" time, so that (4) holds true.

## 3. PATCH SAMPLING

When patch sampling only is used in the sampling scheme, $\alpha=\beta=0$. The observations in this case will typically consist of a record of the continuous up and down history of the system through a number of cycles. It is observed from (4) that in this case the sufficient statistic for the observations is the pair $(X_+, Y_+)$.

### Estimates of Operational Readiness

Define the following random variable:

$$\begin{aligned} I_1(X, Y) &= 1 \text{ if } X_1 > Y_1 \\ &= 0 \text{ otherwise.} \end{aligned} \tag{5}$$

Then

$$E\{I_1(X, Y)\}=\frac{\lambda}{\lambda+\mu}. \tag{6}$$

Since the sufficient statistic $(X_+, Y_+)$ is complete (see [2]), it follows from Blackwell-Rao and Lehmann-Scheffe theorems that the statistic $E\{I_1(X, Y)\,|X_+, Y_+\}$ is the *UMVU* estimate of $p$. Now

$$\begin{aligned} &E\{I_1(X, Y)\,|X_+, Y_+\} \\ &\quad=Pr\{X_1 > Y_1\,|X_+, Y_+\} \\ &\quad=\frac{\Gamma(a)\Gamma(b)\int\int_R f(x_1, y_1, X_+, Y_+)\,dx_1 dy_1}{\mu^a e^{-\mu X_+} X_+^{a-1}\lambda^b e^{-\lambda Y_+} Y_+^{b-1}}, \end{aligned} \tag{7}$$

where

$$R=\{(x_1, y_1): x_1 > y_1, 0 < x_1 < X_+, 0 < y_1 < Y_+\}$$

and

$$f(x_1, y_1, X_+, Y_+)=\frac{\mu^a\lambda^b e^{-\mu X_+-\lambda Y_+}(X_+-x_1)^{a-2}(Y_+-y_1)^{b-2}}{\Gamma(a-1)\Gamma(b-1)}$$

Denoting $X_+/Y_+$ by $S$ and the *UMVU* estimate of $p$ by $T^*(a, b)$ when $a$ up intervals and $b$ down intervals are observed, we obtain from (7) that when $a \geq 2$ and $b \geq 2$,

$$T^*(a, b)=\begin{cases} 1-(a-1)\sum_{j=0}^{b-1}\binom{b-1}{j}\dfrac{S^j(1-S)^{b-1-j}}{a-1+j} & \text{if } S \leq 1 \\ 1-(a-1)\sum_{j=0}^{a-2}\binom{a-2}{j}\dfrac{S^{-(j+1)}(1-S^{-1})^{a-2-j}}{b+j} & \text{if } S > 1; \end{cases} \tag{8}$$

When a = 1 and b ≥ 2, we obtain

$$T^*(1, b)=\begin{cases} 1 & \text{if } X_1 > Y_+ \\ 1-\left(1-\dfrac{X_1}{Y_+}\right)^{b-1} & \text{if } X_1 < Y_+ \ ; \end{cases} \tag{9}$$

When a > 1 and b = 1, we obtain

$$T^*(a, 1)=\begin{cases} 0 & \text{if } X_+ < Y_1 \\ \left(1-\dfrac{Y_1}{X_+}\right)^{a-1} & \text{if } X_+ > Y_1 \ . \end{cases} \tag{10}$$

In particular, we have

$$T^*(2, 2)=\begin{cases} \dfrac{1}{2}\dfrac{X_+}{Y_+} & \text{if } X_+ \leqslant Y_+ \\ 1-\dfrac{1}{2}\dfrac{Y_+}{X_+} & \text{if } X_+ > Y_+ \end{cases}$$

and

$$T^*(3, 3)=\begin{cases} \dfrac{2}{3}\dfrac{X_+}{Y_+}-\dfrac{1}{6}\dfrac{X_+^2}{Y_+^2} & \text{if } X_+ \leqslant Y_+ \\ 1-\dfrac{2}{3}\dfrac{Y_+}{X_+}+\dfrac{1}{6}\dfrac{Y_+^2}{X_+^2} & \text{if } X_+ > Y_+ \ . \end{cases}$$

In Table 1, we give a comparison of the variance of the UMVU-estimates of $p$ with the mean square error of the maximum likelihood estimate $\hat{p}_{a,b}$, where

$$\hat{p}_{a.b}=\frac{X_+}{X_+ + Y_+}$$

for a few small values of $a$ and $b$. The numerical values reported in this table were obtained by performing a Monte-Carlo experiment where a synthetic system realization was observed through a cycle of consecutive up- and down-times. Five hundred such realizations were examined.

TABLE 1. *Comparison of the Variance of the UMVU Estimate of Operational Readiness With Expected Mean Square Error of the Corresponding Maximum Likelihood Estimate (Monte Carlo Estimates)*

| $p$ | $a=b=2$ | | $a=b=3$ | | $a=b=4$ | | $a=b=5$ | |
|---|---|---|---|---|---|---|---|---|
| | *MLE* | *UMVU*[a] | *MLE* | *UMVU* | *MLE* | *UMVU* | *MLE* | *UMVU* |
| 0.1 | 0.018 | 0.017 | 0.011 | 0.0090 | 0.0061 | 0.0053 | 0.0049 | 0.0042 |
| 0.2 | 0.034 | 0.041 | 0.022 | 0.023 | 0.015 | 0.015 | 0.011 | 0.011 |
| 0.3 | 0.043 | 0.062 | 0.031 | 0.038 | 0.023 | 0.027 | 0.020 | 0.022 |
| 0.4 | 0.046 | 0.075 | 0.036 | 0.048 | 0.027 | 0.034 | 0.024 | 0.027 |
| 0.5 | 0.048 | 0.079 | 0.036 | 0.049 | 0.028 | 0.035 | 0.021 | 0.025 |
| 0.6 | 0.049 | 0.075 | 0.033 | 0.044 | 0.027 | 0.033 | 0.022 | 0.026 |
| 0.7 | 0.050 | 0.062 | 0.029 | 0.035 | 0.024 | 0.028 | 0.017 | 0.020 |
| 0.8 | 0.029 | 0.041 | 0.019 | 0.020 | 0.016 | 0.016 | 0.013 | 0.014 |
| 0.9 | 0.021 | 0.017 | 0.0088 | 0.0070 | 0.0064 | 0.0054 | 0.0055 | 0.0048 |

[a] Exact.

The above table shows that with the mean square error criterion, the *UMVU* estimate is better than the maximum likelihood estimate when $p$ is either small or large. Since in reliability applications we shall be mainly concerned with values of $p \geq 0.9$, it appears that in such situations the *UMVU*-estimate will be a better candidate for use.

**Estimate of Operational Reliability**

We assume that $a \geq 2$ and $b \geq 1$. We define the following random variable:

$$I_2(t) = 1 \quad \text{if } X_2 > t \tag{11}$$
$$= 0 \quad \text{otherwise.}$$

Then it is clear that the statistic

$$V = I_1(X, Y) I_2(t) \tag{12}$$

is an unbiased estimator of $\rho(t)$, and that

$$E\{V | X_+, Y_+\}$$

is the unique *UMVU*-estimate of $\rho(t)$. We have

$$E\{V|X_+, Y_+\} = Pr\{X_1 > Y_1; X_2 > t | X_+, Y_+\} = \frac{\Gamma(a)\Gamma(b) \int\int\int_R g(x_1, x_2, y_1)\, dy_1 dx_2 dx_1}{\mu^a \lambda^b e^{-\mu X_+ - \lambda Y_+} X_+^{a-1} Y_+^{b-1}}, \tag{13}$$

where

$$g(x_1, x_2, y_1) = \{\Gamma(b-1)\Gamma(a-2)\}^{-1} \mu^a \lambda^b e^{-\mu X_+ - \lambda Y_+} (Y_+ - y_1)^{b-2} (X_+ - x_1 - x_2)^{a-3}$$

and

$$R = \{(x_1, x_2, y_1);\ 0 < x_1 < X_+,\ 0 < y_1 < Y_+,\ t < x_2 < X_+ - x_1 \text{ and } x_1 > y_1\}.$$

Upon simplifying, we obtain that the *UMVU*-estimator $V^*(a, b, t)$ of $\rho(t)$ is

$$V^*(a,b,t) = \begin{cases} 0 & \text{if } X_+ < t \\ \dfrac{(X_+ - t)^{a-1}}{X_+^{a-1}} - \dfrac{(a-1)}{X_+^{a-1} Y_+^{b-1}} \displaystyle\sum_{j=0}^{b-1} \binom{b-1}{j} \frac{(X_+ - t)^{a-1+j} (Y_+ - X_+ + t)^{b-1-j}}{a-1+j} & \text{if } 0 \leq X_+ - t \leq Y_+ \\ \dfrac{(X_+ - t)^{a-1}}{X_+^{a-1}} - \dfrac{(a-1)}{X_+^{a-1} Y_+^{b-1}} \displaystyle\sum_{j=0}^{a-2} \binom{a-2}{j} \frac{Y_+^{b+j} (X_+ - Y_+ - t)^{a-2-j}}{b+j} & \text{if } X_+ - t > Y_+. \end{cases} \tag{14}$$

## 4. PATCH-SNAPSHOT SAMPLING

In [1] two special cases of sampling plans were considered where both patches and snapshots are used. They are:

*CASE 1:* A system's up and down history is continuously recorded through $k$ initial up and down periods. Thereafter, $m$ rare snapshot observations are made, $r$ of which show the system to be up.

In this case

$$a=b=k,$$

$$\alpha=r, \beta=m-r.$$

*CASE 2:* A system is observed $m$ times at rare intervals, and each time the system state and the remaining time in that state are recorded. Due to the familiar memoryless property of the exponential distribution, the remaining time will also be exponentially distributed, with the parameter appropriate to the state observed at the beginning of the patch. In this case, therefore, if $r$ is the number of snapshots showing the system to be up,

$$a=\alpha=r, \qquad (r=0, 1, 2, \ldots, m)$$

$$b=\beta=m-r.$$

First considering Case 1, we observe from (4) that the sufficient statistic here is the triple $(X_+, Y_+, r)$ which is easily seen to be not *complete*. Therefore, in this case it is not possible to obtain the *UMVU*-estimates with the use of standard methods. Furthermore it appears that in this case there does not exist any *MVU*-estimate which is uniform.

In Case 2, however, the pair $(X_+, Y_+)$ is a complete sufficient statistic and it is therefore possible to obtain *UMVU*-estimates. We illustrate the construction by obtaining the *UMVU*-estimate for $p$. We note that $r$ will have a binomial distribution with parameters $m$ and $p$. Let $P\{r|X_+, Y_+\}$ denote the conditional probability of observing $r$ snapshots in the "up" state given the sufficient statistic $(X_+, Y_+)$. Thus when $X_+>0$ and $Y_+>0$,

$$P\{r|X_+, Y_+\}=\begin{cases} 0 & \text{when } r=0 \text{ or } r=m \\ \dfrac{\dfrac{1}{(r-1)!(m-r-1)!}\binom{m}{r}X_+^{r-1}Y_+^{m-r-1}}{\displaystyle\sum_{r=1}^{m-1}\frac{1}{(r-1)!(m-r-1)!}\binom{m}{r}X_+^{r-1}Y_+^{m-r-1}}, & (r=1, 2, \ldots, m-1) \end{cases} \tag{15}$$

And when $X_+=0$ or $Y_+=0$,

$$P\{r=0|0, Y_+\}=1, \qquad P\{r=m|X_+, 0\}=1. \tag{16}$$

Now since $E\{r\}=mp$, we obtain that

$$U^*(m)=\frac{1}{m}E\{r|X_+, Y_+\} \tag{17}$$

$$=\frac{1}{m}\sum_{r=0}^{m} rP\{r|X_+, Y_+\}$$

is the *UMVU*-estimate of $p$ in this case. When $m=3$ or 4, we obtain upon simplifying (17) that

$$U^*(3)=\begin{cases} 0 & \text{if } X_+=0 \\ \dfrac{2X_+ + Y_+}{3(X_+ + Y_+)} & \text{if } X_+>0,\ Y_+>0 \\ 1 & \text{if } Y_+=0. \end{cases}$$

and

$$U^*(4)=\begin{cases} 0 & \text{if } X_+=0 \\ \dfrac{3X_+^2 + 6X_+Y_+ + Y_+^2}{4X_+^2 + 12X_+Y_+ + 4Y_+^2} & \text{if } X_+>0,\ Y_+>0 \\ 1 & \text{if } Y_+=0. \end{cases}$$

In Table 2, we give a comparison of the variance of $U^*(m)$ with the mean square error of the maximum likelihood estimate $\hat{p}$, where

$$\hat{p}=\frac{\sqrt{X_+}}{\sqrt{X_+}+\sqrt{Y_+}}$$

for $m=3$ and $m=4$. The numerical values reported in this table were obtained by performing a Monte-Carlo experiment where a synthetic system realization was observed using $\lambda=0.1$ and $\mu=\lambda(1-p)/p$. The sampling interval for snapshots was chosen to be $5(\lambda^{-1}+\mu^{-1})$. Five hundred such realizations were examined. These computations were programmed by Mr. R. Fardo of Westinghouse Research Laboratories.

TABLE 2. *Comparison of the Variance of the UMVU Estimate of Operational Readiness with Expected Mean Square Error of the Corresponding Maximum Likelihood Estimate in Case 2 of Patch-Snapshot Sampling (Monte Carlo Estimates)*

| $p$ | $m=3$ | | $m=4$ | |
|---|---|---|---|---|
| | *MLE* | *UMVU* | *MLE* | *UMVU* |
| 0.1 | 0.014 | 0.027 | 0.011 | 0.020 |
| 0.3 | 0.064 | 0.058 | 0.047 | 0.041 |
| 0.5 | 0.087 | 0.067 | 0.062 | 0.047 |
| 0.7 | 0.076 | 0.066 | 0.045 | 0.040 |
| 0.9 | 0.013 | 0.026 | 0.021 | 0.020 |

This table shows, however, that with the mean square criterion, the maximum likelihood estimate is better than the *UMVU* estimate for large values of $p$ in this case. In a similar vein, we can show that the statistic

$$W^*(m,t)=\begin{cases} 0 & \text{if } X_+ \le t \\ \displaystyle\sum_{r=1}^{m-1} \frac{r}{m}\left(\frac{X_+-t}{X_+}\right)^{r-1} P\{r|X_+,Y_+\} & \text{if } X_+>t \text{ and } Y_+>0 \\ \left(\dfrac{X_+-t}{X_+}\right)^{m-1} & \text{if } X_+>t \text{ and } Y_+=0 \end{cases} \tag{18}$$

is the *UMVU*-estimate of $\rho(t)$ in this case.

Finally, we remark that a generalization of the above is possible for systems having more than two states.

## REFERENCES

[1] Gaver, D. P., and M. Mazumdar, "Statistical Estimation in a Problem of System Reliability," Naval Research Logistics Quarterly **14**, 473–488 (1967).

[2] Lehmann, E., *Notes on the Theory of Estimation*, (University of California Press, Berkeley 4, Calif.), Mimeo Notes (reprinted 1962).

* * *

# A TEST FOR THE HYPOTHESIS THAT TWO EXTREME-VALUE SCALE PARAMETERS ARE EQUAL[1]

Nancy R. Mann

Rocketdyne, A Division of North American Rockwell Corporation
Canoga Park, California

**ABSTRACT**

A statistic is determined for testing the hypothesis of equality for scale parameters from two populations, each of which has the first asymptotic distribution of smallest (extreme) values. The probability distribution is derived for this statistic, and critical values are determined and given in tabular form for a one-sided or two-sided alternative, for censored samples of size $n_1$ and $n_2$, $n_1 = 2, 3, \ldots, 6$, $n_2 = 2, 3, \ldots, 6$. The power function of the test for certain alternatives is also calculated and listed in each case considered.

## INTRODUCTION

This paper deals with a situation in which independent random samples of size $n_2$ and $n_2$ from wo populations of items are subjected to life test under identical environmental conditions until $r \leq n_1$ and $s \leq n_2$ of the items of the respective samples have failed. It is assumed that each of the sets of ordered observations, $X_{1.1} \leq X_{2.1} \leq \ldots \leq X_{r.1}$ and $X_{1.2} \leq X_{2.2} \leq \ldots \leq X_{s.2}$, of logarithms of failure times resulting from the life tests is from a population consisting of variates having the first asymptotic distribution of smallest (extreme) values; and that the scale parameters of the two distributions are $b_1$ and $b_2$, respectively. An equivalent assumption is that the two populations from which the sample failure times are slected have two-parameter Weibull distributions with shape parameters $b_1$ and $b_2$. It is also assumed that one desires to test the hypothesis $H: b_1 = b_2$ versus the alternative $K_1: b_1 < b_2$ or the alternative $K_2: b_1 \neq b_2$, where for the one-sided alternative the samples have been assigned the subscript 1 or 2 according to some criterion other than the size of the estimates of $b_1$ and $b_2$ obtained from the observed sample values.

It may sometimes be assumed on the basis of prior information concerning the type of item tested that $b_1$ and $b_2$ are independent of stress levels maintained during the two life tests, as in [8]. In such a case, one can test for a difference in $b_1$ and $b_2$ (a difference in the variability of the material in the items under life test) whether or not the two stress levels maintained during the two life tests are identical. If the assumption that the scale parameters are independent of level of stress cannot be made, then the stress levels as well as other environmental conditions must be the same in the two life tests.

The probability density function of the first asymptotic distribution of smallest values (or the extreme-value distribution) is given by:

$$f(x) = (1/b) \exp[(x-u)/b] \exp\{-\exp[(x-u)/b]\}, \quad b > 0. \tag{1}$$

The standard deviation of the distribution is $\pi b/\sqrt{6}$ and the mode occurs at $x = u$. Under the assumption that $X_1 \leq X_2 \leq \ldots \leq X_r$ represent the first $r$ of $n$ ordered sample failure times randomly selected

[1] This research was supported by the Aerospace Research Laboratories, Office of Aerospace Research, United States Air Force, under Contract AF33(615)-2818.

from a population of variates with density given by (1), $LC_{l,m}(\alpha)$ was derived and $l$ and $m$ determined in [4] such that $b_{l,m}=(X_m-X_l)/LC_{l,m}(\alpha)$, $1<l<m\leq r$, has smallest expected squared deviation from $b$ (smallest risk) among $(1-\alpha)$-level upper confidence bounds based on any two of $X_1, X_2, \ldots, X_r$, and independent of $u$. Values of $l$, $m$, and $LC_{l,m}(\alpha)$ were calculated and appear in Table 1 of [7] and of [7] for $n=2, 3, \ldots, 12$, $m=2, 3, \ldots, n$, $\alpha=0.01$, 0.05, and 0.10 and in Table B.1 of [4] through $n=22$. The bound $b_{l,m}$ with smallest risk among invariant two-order-statistic bounds (a criterion suggested by [1]) also represents essentially, though not precisely, the uniformly most accurate of these bounds at confidence level $1-\alpha$ (see [7]).

For small $r$, one could improve upon such bounds only slightly, as described in [7] and [6], by using an appropriate linear combination of all $r$ observations, instead of only $X_l$ and $X_m$, (see, for example, [5]) and values comparable to $LC_{l,m}(\alpha)$ obtained by Monte Carlo methods. Values for obtaining bounds based on this Monte Carlo procedure have not been determined for any sample size, to the knowledge of this author. In the following, an exact test of $H:b_1=b_2$ versus $K_1:b_1<b_2$ or $K_2:b_1\neq b_2$, for which critical levels can be otained analytically, is derived. It is based on two ordered observations from each of the two samples.

## DERIVATION OF A TEST FOR H

Consider the random variable

$$Z=\frac{X_{q_1,1}-X_{p_1,1}}{X_{q_2,2}-X_{p_2,2}},$$

where $p_1$ and $q_1$ and $p_2$ and $q_2$ are the $p$ and $q$, $p<q$, such that $C(X_q-X_p)$ (as given in Table III of [7]) has smallest mean squared error among two-order-statistic invariant estimators of $b$ when the first $r$ of $n_1$ and the first $s$ of $n_2$ sample failure times are observed. Let $Y_{l_i,i}$ equal $(X_{l_i,i}-u_i)|b_i$, the $l_i$th reduced order statistic, with parameter-free distribution, from the $i$th sample of size $n_i$, $l_i=1, 2, \ldots, n_i$, and let $V_i$ be equal to $b_i(Y_{q_i,i}-Y_{p_i,i})=(X_{q_i,i}-X_{p_i,i})$ $1\leq p_i<q_i<n_i$, $i=1, 2, \ldots$.

The density function of $V_i$ is given by

$$(2)\qquad f(v_i)=\frac{n_i!}{b_i(p_i-1)!(q_i-p_i-1)!(n_i-q_i)!}\sum_{k=0}^{p_i-1}\sum_{m=0}^{q_i-p_i-1}\frac{(-1)^{k+m}\binom{p_i-1}{k}\binom{q_i-p_i-1}{m}\exp\left(\frac{v_i}{b_i}\right)}{\left[(q_i-p_i-m+k)+(n_i-q_i+m+1)\exp\left(\frac{v_i}{b_i}\right)\right]^2}.$$

Using (2), one can show that $Q_0(\alpha)$, the $100\,\alpha$ percent point of the distribution of

$$Q=Z/(b_1/b_2)=(Y_{q_1,1}-Y_{p_1,1})/(Y_{q_2,2}-Y_{p_2,2})$$

is defined implicitly by

$$P[Q<Q_0(\alpha)]\equiv$$

$$(3)\qquad 1-\int_1^{\infty}\sum_{j=0}^{p_1-1}\sum_{k=0}^{p_2-1}\sum_{l=0}^{q_1-p_1-1}\sum_{m=0}^{q_2-p_2-1}\frac{[C_1(j,k,l,m)/C_3(l)]dx}{[\{C_4(m,k)+C_5(m)x\}^2\{C_2(l,j)+C_3(l)x^{Q_0(\alpha)}\}]}=\alpha.$$

Here

$$C_1(j,k,l,m)=\frac{n_1!n_2!(-1)^{j+k+l+m}\binom{p_1-1}{j}\binom{p_2-1}{k}\binom{q_1-p_1-1}{l}\binom{q_2-p_2-1}{m}}{(p_1-1)!(p_2-1)!(q_1-p_1-1)!(q_2-p_2-1)!(n_1-q_1)!(n_2-q_2)!}$$

$$C_2(l, j) = q_1 - p_1 - l + j$$

$$C_3(l) = n_1 - q_1 + l + 1$$

$$C_4(m, k) = q_2 - p_2 - m + k$$

and

$$C_5(m) = n_2 - q_2 + m + 1.$$

Moreover, when $b_1 = b_2$, $Z = b_1(Y_{q1,1} - Y_{p1,1})/[b_2(Y_{q2,2} - Y_{p2,2})]$ has the same distribution as $Q$. Thus, $P[Z < Q_0(\alpha)]$ is equal to $\alpha$ when $b_1 = b_2$, and one will reject the hypothesis $H$: $b_1 = b_2$ with probability $\alpha$ when $b_1$ is equal to $b_2$ if he rejects $H$ whenever $Z$ is less than $Q_0(\alpha)$. Also, since

$$P[Z/(b_1/b_2) \geq Q_0(\alpha)] = P[b_1/b_2 \leq Z/Q_0(\alpha)] = 1 - \alpha,$$

$Z/Q_0(\alpha)$ is an upper confidence bound for $b_1/b_2$ at confidence level $1 - \alpha$. Note, too, that $P[b_2/b_1 \geq Q_0/Z]$. In obtaining one-sided confidence bounds, one would usually, though not necessarily, identify the subscript $l$ with the Weibull population thought to have the smaller shape parameter.

For obtaining two-sided confidence bounds, one can make use of the fact that

$$P[Q > Q_1] = P[1/Q < 1/Q_1] = P[Z'/(b_2/b_1) < 1/Q_1], \tag{4}$$

where $Z' = 1/Z$. If $1/Q_1$ is equal to $Q_0'(\alpha)$, where

$$P\left[\frac{1}{Q} < Q_0'(\alpha)\right] = P\left[Q > \frac{1}{Q_0'(\alpha)}\right] = \alpha;$$

then the probability given by (4) is equal to $\alpha$. Thus,

$$P[Q_0(\alpha) \leq Z/(b_1/b_2) \leq 1/Q'_0(\alpha)] = P[ZQ_0'(\alpha) \leq b_1/b_2 \leq Z/Q_0(\alpha)] \tag{5}$$

is equal to $1 - 2\alpha$, and two sided bounds are determined for $b_1/b_2$ at level $1 - 2\alpha$. From (5) it can be seen that a test for equality of $b_1$ and $b_2$ versus the alternative $K_2$:$b_1 \neq b_2$ rejects $K_2$ at significance level $2\alpha$ when $Z$ is less than $Q_0(\alpha)$ or greater than $1/Q_0'(\alpha)$.

## CALCULATION OF $Q_0(\alpha)$

Values of $Q_0 = Q_0(\alpha)$ have been calculated, using numerical methods of integration for all combinations of $n_1$, $n_2$, $r$, and $s$, $2 \leq r \leq n_1 \leq 6$, and $2 \leq s \leq n_2 \leq 6$, $\alpha = 0.10$, and appear in Table 1. Also calculated were values of the power function for $b_1/b_2$ equal to $\frac{1}{2}$ and $\frac{1}{4}$. These also appear in Table 1. The power calculation was accomplished by integrating the density function of $Z = (b_1/b_2)Q$ from 0 to $Q_0(\alpha)$. It is interesting to note that for $n_1 = n_2 = r = s = 2$, $b_1/b_2 = \frac{1}{4}$, the power function is equal to 0.347, while for testing $b = b_0$ versus $b < b_0$ as in [3], using both of two observations, the power function is equal to 0.381 for $b/b_0 = \frac{1}{4}$. (This test of $b = b_0$ is shown in [3] to be uniformly most powerful among invariant tests when only two failures are observed.) For $n = r = 6$ and $n_1 = n_2 = r = s = 6$, the corresponding values are 0.858 for $b_1/b_2 = \frac{1}{4}$ and 0.984 for $b/b_0 = \frac{1}{4}$. For this combination of $n_1$, $n_2$, $r$, and $s$, the value calculated is based on $Y_{1,1}$, $Y_{6,1}$, $Y_{1,2}$, and $Y_{6,2}$ rather than on $Y_{2,1}$, $Y_{6,1}$, $Y_{2,2}$, and $Y_{6,2}$, the combination corresponding to the best linear invariant estimators of $b_1$ and $b_2$ based on two ordered observations. In this way, the number of terms in (3) was reduced from 64 to 25, while the difference in mean squared error between the two corresponding sets of estimators is only 0.003 multiplied by the scale parameter, $b_1$ or $b_2$.

The calculated values of the power function for the test based on $Z$ can be compared with those of other tests if and when power-function values associated with other test statistics become available. From (2) and (3), it can be determined that the derivative with respect to the ratio $b_1/b_2$ of the power function for the one-sided test is equal to the density function for $Q$, evaluated at $(b_2/b_1)\,Q_0$, multiplied by a negative constant. Both the one- and two-sided tests are therefore unbiased.

## ACCURACY OF THE TABLES

The values of $Q_0(\alpha)$ were obtained by a modified Newton-Raphson iteration process in which after the $i$th set of three successive values of $Q$, $Q_{k-1,i}$, $Q_{k,i}$, $Q_{k+1,i}$, had been calculated by the Newton-Raphson procedure, the value tried for $Q_0(\alpha)$ after $Q_{k,i}$ was not $Q_{k+1,i}$, but rather

$$Q'_{k+1,i}=Q_{k+1,i}-(Q_{k+1,i}-Q_{k,i})^2/(Q_{k+1,i}-2Q_{k,i}+Q_{k-1,i}).$$

The first guess was 0.5 and the process was terminated whenever the first six significant figures of any value of $Q$ agreed with those of the value used in the preceding iteration.

A three-point Gaussian quadrature formula was used for numerical integration of the expression given in (3) and its derivative with respect to $Q_0(\alpha)$. A process was utilized whereby the size of the interval for the numerical integration was dependent upon the size of the proportion of the area under the curve up to the point of integration, added by integrating over the previous interval. Fortran IV built-in double precision was used throughout the main program and all sub-routines.

An attempt was made to determine some notion of how much accuracy was lost by the loss of significant figures in adding the $p_1p_2(q_1-p_1)(q_2-p_2)$ terms, the signs of which tend to alternate, in the expression for the 100 $\alpha$ percent point of the distribution of $Q$. To accomplish this, $Q_0(\alpha)$ in (3) was set equal to zero in each case considered and the numerical integration performed for each term. The amount that the sum of the terms in the integral deviated from 1 gave an indication of the increase in loss of accuracy attributable to the addition of successively larger numbers of terms. When the worst case $p_1=p_2=1$, $q_1=q_2=6$, (25 terms), was encountered, there appeared to have been only a very slight loss of accuracy so that the last one of the six significant figures determined is in doubt.

## REFERENCES

[1] Harter, H. Leon, "Criteria for Best Substitute Interval Estimators, with an Application to the Normal Distribution," J. Amer. Statist. Assoc., **59**, 1133–1140 (1964).

[2] Mann, N. R., "Optimum Estimates of Parameters of Continuous Distributions," Rocketdyne Research Report 63–41. Rocketdyne, Canoga Park, California (1963).

[3] Mann, N. R., "Point and Interval Estimates for Reliability Parameters When Failure Times Have the Two-parameter Weibull Distribution," (Unpublished doctoral dissertation), University of California at Los Angeles (1965).

[4] Mann, Nancy R., "Results on Location and Scale Parameter Estimation with Application to the Extreme-Value Distribution," Aerospace Research Laboratories Report ARL 67–0023, Office of Aerospace Research, United States Air Force, Wright-Patterson Air Force Base, Ohio (1967).

[5] Mann, Nancy R., "Tables for Obtaining the Best Linear Invariant Estimates of Parameters of the Weibull Distribution," Technometrics, **9**, 629–645 (1967).

[6] Mann, Nancy R., "Point and Interval Estimation Procedures for the Two-Parameter Weibull and Extreme-Value Distributions," Technometrics, **10**, 231–256 (1968).

[7] Mann, Nancy R., Efficient Estimators and Exact Confidence Bounds for Weibull Parameters Based on a Few Ordered Observations. To appear in Technometrics.

[8] Saunders, S. C., "On the Determination of a Safe Life for Classes of Distributions Classified by Failure Rate," Technometrics, **10**, 361–377 (1968).

TABLE 1.—*Critical Values for .10-Level One-Sided Test for Equality of Extreme-Value Scale Parameters*

| $n_1$ | $r$ | $p_1$ | $q_1$ | $n_2$ | $s$ | $p_2$ | $q_2$ | $Q_0$ | Power for | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | $b_1/b_2 = .25$ | $b_1/b_2 = .50$ |
| 2 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | .146 | .347 | .193 |
| 2 | 2 | 1 | 2 | 3 | 2 | 1 | 2 | .167 | .342 | .192 |
| 2 | 2 | 1 | 2 | 3 | 3 | 1 | 3 | .097 | .363 | .196 |
| 2 | 2 | 1 | 2 | 4 | 2 | 1 | 2 | .177 | .340 | .192 |
| 2 | 2 | 1 | 2 | 4 | 3 | 1 | 3 | .109 | .360 | .195 |
| 2 | 2 | 1 | 2 | 4 | 4 | 1 | 4 | .079 | .368 | .196 |
| 2 | 2 | 1 | 2 | 5 | 2 | 1 | 2 | .182 | .339 | .191 |
| 2 | 2 | 1 | 2 | 5 | 3 | 1 | 3 | .115 | .358 | .195 |
| 2 | 2 | 1 | 2 | 5 | 4 | 1 | 4 | .088 | .366 | .196 |
| 2 | 2 | 1 | 2 | 5 | 5 | 1 | 5 | .070 | .371 | .197 |
| 2 | 2 | 1 | 2 | 6 | 2 | 1 | 2 | .186 | .338 | .191 |
| 2 | 2 | 1 | 2 | 6 | 3 | 1 | 3 | .118 | .358 | .195 |
| 2 | 2 | 1 | 2 | 6 | 4 | 1 | 4 | .092 | .365 | .196 |
| 2 | 2 | 1 | 2 | 6 | 5 | 1 | 5 | .077 | .369 | .196 |
| 2 | 2 | 1 | 2 | 6 | 6 | 1 | 6 | .064 | .372 | .197 |
| 3 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | .114 | .332 | .189 |
| 3 | 2 | 1 | 2 | 3 | 2 | 1 | 2 | .130 | .328 | .188 |
| 3 | 2 | 1 | 2 | 3 | 3 | 1 | 3 | .075 | .346 | .191 |
| 3 | 2 | 1 | 2 | 4 | 2 | 1 | 2 | .138 | .327 | .187 |
| 3 | 2 | 1 | 2 | 4 | 3 | 1 | 3 | .084 | .344 | .191 |
| 3 | 2 | 1 | 2 | 4 | 4 | 1 | 4 | .061 | .351 | .192 |
| 3 | 2 | 1 | 2 | 5 | 2 | 1 | 2 | .142 | .326 | .187 |
| 3 | 2 | 1 | 2 | 5 | 3 | 1 | 3 | .089 | .343 | .191 |
| 3 | 2 | 1 | 2 | 5 | 4 | 1 | 4 | .068 | .349 | .192 |
| 3 | 2 | 1 | 2 | 5 | 5 | 1 | 5 | .054 | .353 | .192 |
| 3 | 2 | 1 | 2 | 6 | 2 | 1 | 2 | .145 | .325 | .187 |
| 3 | 2 | 1 | 2 | 6 | 3 | 1 | 3 | .092 | .342 | .190 |
| 3 | 2 | 1 | 2 | 6 | 4 | 1 | 4 | .071 | .348 | .192 |
| 3 | 2 | 1 | 2 | 6 | 5 | 1 | 5 | .059 | .352 | .192 |
| 3 | 2 | 1 | 2 | 6 | 6 | 1 | 6 | .049 | .354 | .193 |
| 3 | 3 | 1 | 3 | 2 | 2 | 1 | 2 | .420 | .497 | .262 |

TABLE 1.—*Critical Values for .10-Level One-Sided Test for Equality of Extreme-Value Scale Parameters*

| $n_1$ | $r$ | $p_1$ | $q_1$ | $n_2$ | $s$ | $p_2$ | $q_2$ | $Q_0$ | Power for | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | $b_1/b_2=.25$ | $b_1/b_2=.50$ |
| 3 | 3 | 1 | 3 | 3 | 2 | 1 | 2 | .472 | .479 | .256 |
| 3 | 3 | 1 | 3 | 3 | 3 | 1 | 3 | .300 | .580 | .286 |
| 3 | 3 | 1 | 3 | 4 | 2 | 1 | 2 | .496 | .471 | .253 |
| 3 | 3 | 1 | 3 | 4 | 3 | 1 | 3 | .332 | .564 | .281 |
| 3 | 3 | 1 | 3 | 4 | 4 | 1 | 4 | .253 | .618 | .297 |
| 3 | 3 | 1 | 3 | 5 | 2 | 1 | 2 | .510 | .467 | .251 |
| 3 | 3 | 1 | 3 | 5 | 3 | 1 | 3 | .348 | .556 | .278 |
| 3 | 3 | 1 | 3 | 5 | 4 | 1 | 4 | .277 | .605 | .292 |
| 3 | 3 | 1 | 3 | 5 | 5 | 1 | 5 | .226 | .640 | .303 |
| 3 | 3 | 1 | 3 | 6 | 2 | 1 | 2 | .519 | .464 | .250 |
| 3 | 3 | 1 | 3 | 6 | 3 | 1 | 3 | .358 | .551 | .277 |
| 3 | 3 | 1 | 3 | 6 | 4 | 1 | 4 | .290 | .597 | .290 |
| 3 | 3 | 1 | 3 | 6 | 5 | 1 | 5 | .246 | .629 | .299 |
| 3 | 3 | 1 | 3 | 6 | 6 | 1 | 6 | .209 | .655 | .306 |
| 4 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | .102 | .327 | .187 |
| 4 | 2 | 1 | 2 | 3 | 2 | 1 | 2 | .117 | .323 | .186 |
| 4 | 2 | 1 | 2 | 3 | 3 | 1 | 3 | .067 | .340 | .190 |
| 4 | 2 | 1 | 2 | 4 | 2 | 1 | 2 | .124 | .322 | .186 |
| 4 | 2 | 1 | 2 | 4 | 3 | 1 | 3 | .076 | .338 | .189 |
| 4 | 2 | 1 | 2 | 4 | 4 | 1 | 4 | .055 | .345 | .191 |
| 4 | 2 | 1 | 2 | 5 | 2 | 1 | 2 | .128 | .321 | .186 |
| 4 | 2 | 1 | 2 | 5 | 3 | 1 | 3 | .080 | .337 | .189 |
| 4 | 2 | 1 | 2 | 5 | 4 | 1 | 4 | .061 | .343 | .190 |
| 4 | 2 | 1 | 2 | 5 | 5 | 1 | 5 | .048 | .347 | .191 |
| 4 | 2 | 1 | 2 | 6 | 2 | 1 | 2 | .131 | .320 | .186 |
| 4 | 2 | 1 | 2 | 6 | 3 | 1 | 3 | .082 | .336 | .189 |
| 4 | 2 | 1 | 2 | 6 | 4 | 1 | 4 | .064 | .343 | .190 |
| 4 | 2 | 1 | 2 | 6 | 5 | 1 | 5 | .053 | .346 | .191 |
| 4 | 2 | 1 | 2 | 6 | 6 | 1 | 6 | .044 | .348 | .191 |
| 4 | 3 | 1 | 3 | 2 | 2 | 1 | 2 | .350 | .480 | .254 |
| 4 | 3 | 1 | 3 | 3 | 2 | 1 | 2 | .395 | .464 | .248 |
| 4 | 3 | 1 | 3 | 3 | 3 | 1 | 3 | .248 | .554 | .275 |
| 4 | 3 | 1 | 3 | 4 | 2 | 1 | 2 | .415 | .457 | .246 |
| 4 | 3 | 1 | 3 | 4 | 3 | 1 | 3 | .275 | .540 | .270 |
| 4 | 3 | 1 | 3 | 4 | 4 | 1 | 4 | .208 | .587 | .283 |
| 4 | 3 | 1 | 3 | 5 | 2 | 1 | 2 | .427 | .453 | .244 |
| 4 | 3 | 1 | 3 | 5 | 3 | 1 | 3 | .289 | .533 | .268 |
| 4 | 3 | 1 | 3 | 5 | 4 | 1 | 4 | .229 | .575 | .280 |
| 4 | 3 | 1 | 3 | 5 | 5 | 1 | 5 | .186 | .606 | .288 |

TABLE 1.—*Critical Values for .10-Level One-Sided Test for Equality of Extreme-Value Scale Parameters*

| $n_1$ | $r$ | $p_1$ | $q_1$ | $n_2$ | $s$ | $p_2$ | $q_2$ | $Q_0$ | Power for | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | $b_1/b_2=.25$ | $b_1/b_2=.50$ |
| 4 | 3 | 1 | 3 | 6 | 2 | 1 | 2 | .435 | .451 | .244 |
| 4 | 3 | 1 | 3 | 6 | 3 | 1 | 3 | .297 | .528 | .267 |
| 4 | 3 | 1 | 3 | 6 | 4 | 1 | 4 | .240 | .569 | .278 |
| 4 | 3 | 1 | 3 | 6 | 5 | 1 | 5 | .203 | .596 | .286 |
| 4 | 3 | 1 | 3 | 6 | 6 | 1 | 6 | .171 | .618 | .291 |
| 4 | 4 | 1 | 4 | 2 | 2 | 1 | 2 | .610 | .547 | .293 |
| 4 | 4 | 1 | 4 | 3 | 2 | 1 | 2 | .682 | .523 | .282 |
| 4 | 4 | 1 | 4 | 3 | 3 | 1 | 3 | .448 | .662 | .335 |
| 4 | 4 | 1 | 4 | 4 | 2 | 1 | 2 | .715 | .513 | .278 |
| 4 | 4 | 1 | 4 | 4 | 3 | 1 | 3 | .493 | .639 | .326 |
| 4 | 4 | 1 | 4 | 4 | 4 | 1 | 4 | .382 | .718 | .357 |
| 4 | 4 | 1 | 4 | 5 | 2 | 1 | 2 | .735 | .508 | .276 |
| 4 | 4 | 1 | 4 | 5 | 3 | 1 | 3 | .516 | .628 | .321 |
| 4 | 4 | 1 | 4 | 5 | 4 | 1 | 4 | .417 | .698 | .348 |
| 4 | 4 | 1 | 4 | 5 | 5 | 1 | 5 | .344 | .751 | .370 |
| 4 | 4 | 1 | 4 | 6 | 2 | 1 | 2 | .747 | .504 | .274 |
| 4 | 4 | 1 | 4 | 6 | 3 | 1 | 3 | .530 | .622 | .318 |
| 4 | 4 | 1 | 4 | 6 | 4 | 1 | 4 | .436 | .688 | .344 |
| 4 | 4 | 1 | 4 | 6 | 5 | 1 | 5 | .374 | .734 | .362 |
| 4 | 4 | 1 | 4 | 6 | 6 | 1 | 6 | .319 | .773 | .379 |
| 5 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | .097 | .324 | .186 |
| 5 | 2 | 1 | 2 | 3 | 2 | 1 | 2 | .111 | .321 | .185 |
| 5 | 2 | 1 | 2 | 3 | 3 | 1 | 3 | .064 | .338 | .189 |
| 5 | 2 | 1 | 2 | 4 | 2 | 1 | 2 | .117 | .319 | .185 |
| 5 | 2 | 1 | 2 | 4 | 3 | 1 | 3 | .071 | .335 | .189 |
| 5 | 2 | 1 | 2 | 4 | 4 | 1 | 4 | .052 | .342 | .190 |
| 5 | 2 | 1 | 2 | 5 | 2 | 1 | 2 | .121 | .318 | .185 |
| 5 | 2 | 1 | 2 | 5 | 3 | 1 | 3 | .075 | .334 | .188 |
| 5 | 2 | 1 | 2 | 5 | 4 | 1 | 4 | .057 | .340 | .190 |
| 5 | 2 | 1 | 2 | 5 | 5 | 1 | 5 | .046 | .344 | .190 |
| 5 | 2 | 1 | 2 | 6 | 2 | 1 | 2 | .124 | .318 | .185 |
| 5 | 2 | 1 | 2 | 6 | 3 | 1 | 3 | .078 | .334 | .188 |
| 5 | 2 | 1 | 2 | 6 | 4 | 1 | 4 | .060 | .340 | .189 |
| 5 | 2 | 1 | 2 | 6 | 5 | 1 | 5 | .050 | .343 | .190 |
| 5 | 2 | 1 | 2 | 6 | 6 | 1 | 6 | .042 | .345 | .191 |
| 5 | 3 | 1 | 3 | 2 | 2 | 1 | 2 | .321 | .472 | .250 |
| 5 | 3 | 1 | 3 | 3 | 2 | 1 | 2 | .362 | .457 | .245 |
| 5 | 3 | 1 | 3 | 3 | 3 | 1 | 3 | .227 | .542 | .269 |

TABLE 1.—*Critical Values for .10-Level One-Sided Test for Equality of Extreme-Value Scale Parameters*

| $n_1$ | $r$ | $p_1$ | $q_1$ | $n_2$ | $s$ | $p_2$ | $q_2$ | $Q_0$ | Power for | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | $b_1/b_2=.25$ | $b_1/b_2=.50$ |
| 5 | 3 | 1 | 3 | 4 | 2 | 1 | 2 | .381 | .450 | .243 |
| 5 | 3 | 1 | 3 | 4 | 3 | 1 | 3 | .251 | .528 | .265 |
| 5 | 3 | 1 | 3 | 4 | 4 | 1 | 4 | .190 | .572 | .278 |
| 5 | 3 | 1 | 3 | 5 | 2 | 1 | 2 | .392 | .446 | .241 |
| 5 | 3 | 1 | 3 | 5 | 3 | 1 | 3 | .264 | .522 | .263 |
| 5 | 3 | 1 | 3 | 5 | 4 | 1 | 4 | .209 | .561 | .274 |
| 5 | 3 | 1 | 3 | 5 | 5 | 1 | 5 | .169 | .590 | .282 |
| 5 | 3 | 1 | 3 | 6 | 2 | 1 | 2 | .399 | .444 | .241 |
| 5 | 3 | 1 | 3 | 6 | 3 | 1 | 3 | .272 | .518 | .262 |
| 5 | 3 | 1 | 3 | 6 | 4 | 1 | 4 | .219 | .555 | .273 |
| 5 | 3 | 1 | 3 | 6 | 5 | 1 | 5 | .185 | .581 | .280 |
| 5 | 3 | 1 | 3 | 6 | 6 | 1 | 6 | .156 | .601 | .285 |
| 5 | 4 | 1 | 4 | 2 | 2 | 1 | 2 | .527 | .535 | .286 |
| 5 | 4 | 1 | 4 | 3 | 2 | 1 | 2 | .590 | .512 | .576 |
| 5 | 4 | 1 | 4 | 3 | 3 | 1 | 3 | .384 | .642 | .324 |
| 5 | 4 | 1 | 4 | 4 | 2 | 1 | 2 | .619 | .503 | .272 |
| 5 | 4 | 1 | 4 | 4 | 3 | 1 | 3 | .424 | .621 | .315 |
| 5 | 4 | 1 | 4 | 4 | 4 | 1 | 4 | .327 | .693 | .342 |
| 5 | 4 | 1 | 4 | 5 | 2 | 1 | 2 | .636 | .498 | .270 |
| 5 | 4 | 1 | 4 | 5 | 3 | 1 | 3 | .444 | .610 | .311 |
| 5 | 4 | 1 | 4 | 5 | 4 | 1 | 4 | .357 | .675 | .335 |
| 5 | 4 | 1 | 4 | 5 | 5 | 1 | 5 | .294 | .724 | .354 |
| 5 | 4 | 1 | 4 | 6 | 2 | 1 | 2 | .647 | .495 | .268 |
| 5 | 4 | 1 | 4 | 6 | 3 | 1 | 3 | .456 | .604 | .308 |
| 5 | 4 | 1 | 4 | 6 | 4 | 1 | 4 | .373 | .666 | .331 |
| 5 | 4 | 1 | 4 | 6 | 5 | 1 | 5 | .319 | .708 | .347 |
| 5 | 4 | 1 | 4 | 6 | 6 | 1 | 6 | .272 | .744 | .361 |
| 5 | 5 | 1 | 5 | 2 | 2 | 1 | 2 | .749 | .571 | .309 |
| 5 | 5 | 1 | 5 | 3 | 2 | 1 | 2 | .835 | .544 | .296 |
| 5 | 5 | 1 | 5 | 3 | 3 | 1 | 3 | .558 | .701 | .364 |
| 5 | 5 | 1 | 5 | 4 | 2 | 1 | 2 | .875 | .533 | .291 |
| 5 | 5 | 1 | 5 | 4 | 3 | 1 | 3 | .613 | .675 | .350 |
| 5 | 5 | 1 | 5 | 4 | 4 | 1 | 4 | .479 | .765 | .393 |
| 5 | 5 | 1 | 5 | 5 | 2 | 1 | 2 | .898 | .527 | .288 |
| 5 | 5 | 1 | 5 | 5 | 3 | 1 | 3 | .640 | .663 | .344 |
| 5 | 5 | 1 | 5 | 5 | 4 | 1 | 4 | .522 | .743 | .381 |
| 5 | 5 | 1 | 5 | 5 | 5 | 1 | 5 | .434 | .803 | .412 |
| 5 | 5 | 1 | 5 | 6 | 2 | 1 | 2 | .913 | .523 | .286 |
| 5 | 5 | 1 | 5 | 6 | 3 | 1 | 3 | .657 | .655 | .341 |
| 5 | 5 | 1 | 5 | 6 | 4 | 1 | 4 | .544 | .731 | .375 |

TABLE 1.—*Critical Values for .10-Level One-Sided Test for Equality of Extreme-Value Scale Parameters*

| $n_1$ | $r$ | $p_1$ | $q_1$ | $n_2$ | $s$ | $p_2$ | $q_2$ | $Q_0$ | Power for | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | $b_1/b_2=.25$ | $b_1/b_2=.50$ |
| 5 | 5 | 1 | 5 | 6 | 5 | 1 | 5 | .470 | .785 | .401 |
| 5 | 5 | 1 | 5 | 6 | 6 | 1 | 6 | .403 | .828 | .426 |
| 6 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | .093 | .323 | .186 |
| 6 | 2 | 1 | 2 | 3 | 2 | 1 | 2 | .107 | .319 | .185 |
| 6 | 2 | 1 | 2 | 3 | 3 | 1 | 3 | .061 | .336 | .189 |
| 6 | 2 | 1 | 2 | 4 | 2 | 1 | 2 | .113 | .318 | .185 |
| 6 | 2 | 1 | 2 | 4 | 3 | 1 | 3 | .069 | .334 | .188 |
| 6 | 2 | 1 | 2 | 4 | 4 | 1 | 4 | .050 | .340 | .189 |
| 6 | 2 | 1 | 2 | 5 | 2 | 1 | 2 | .117 | .317 | .184 |
| 6 | 2 | 1 | 2 | 5 | 3 | 1 | 3 | .073 | .333 | .188 |
| 6 | 2 | 1 | 2 | 5 | 4 | 1 | 4 | .055 | .339 | .189 |
| 6 | 2 | 1 | 2 | 5 | 5 | 1 | 5 | .044 | .342 | .190 |
| 6 | 2 | 1 | 2 | 6 | 2 | 1 | 2 | .119 | .316 | .184 |
| 6 | 2 | 1 | 2 | 6 | 3 | 1 | 3 | .075 | .332 | .188 |
| 6 | 2 | 1 | 2 | 6 | 4 | 1 | 4 | .058 | .338 | .189 |
| 6 | 2 | 1 | 2 | 6 | 5 | 1 | 5 | .048 | .341 | .190 |
| 6 | 2 | 1 | 2 | 6 | 6 | 1 | 6 | .040 | .344 | .190 |
| 6 | 3 | 1 | 3 | 2 | 2 | 1 | 2 | .305 | .468 | .248 |
| 6 | 3 | 1 | 3 | 3 | 2 | 1 | 2 | .344 | .452 | .243 |
| 6 | 3 | 1 | 3 | 3 | 3 | 1 | 3 | .214 | .534 | .266 |
| 6 | 3 | 1 | 3 | 4 | 2 | 1 | 2 | .362 | .446 | .241 |
| 6 | 3 | 1 | 3 | 4 | 3 | 1 | 3 | .238 | .521 | .263 |
| 6 | 3 | 1 | 3 | 4 | 4 | 1 | 4 | .180 | .564 | .274 |
| 6 | 3 | 1 | 3 | 5 | 2 | 1 | 2 | .372 | .442 | .239 |
| 6 | 3 | 1 | 3 | 5 | 3 | 1 | 3 | .250 | .515 | .261 |
| 6 | 3 | 1 | 3 | 5 | 4 | 1 | 4 | .197 | .553 | .271 |
| 6 | 3 | 1 | 3 | 5 | 5 | 1 | 5 | .160 | .580 | .279 |
| 6 | 3 | 1 | 3 | 6 | 2 | 1 | 2 | .379 | .440 | .239 |
| 6 | 3 | 1 | 3 | 6 | 3 | 1 | 3 | .257 | .511 | .260 |
| 6 | 3 | 1 | 3 | 6 | 4 | 1 | 4 | .207 | .548 | .270 |
| 6 | 3 | 1 | 3 | 6 | 5 | 1 | 5 | .175 | .572 | .276 |
| 6 | 3 | 1 | 3 | 6 | 6 | 1 | 6 | .147 | .591 | .281 |
| 6 | 4 | 1 | 4 | 2 | 2 | 1 | 2 | .489 | .528 | .282 |
| 6 | 4 | 1 | 4 | 3 | 2 | 1 | 2 | .548 | .506 | .273 |
| 6 | 4 | 1 | 4 | 3 | 3 | 1 | 3 | .355 | .631 | .317 |
| 6 | 4 | 1 | 4 | 4 | 2 | 1 | 2 | .575 | .497 | .269 |
| 6 | 4 | 1 | 4 | 4 | 3 | 1 | 3 | .392 | .611 | .309 |
| 6 | 4 | 1 | 4 | 4 | 4 | 1 | 4 | .302 | .680 | .335 |

TABLE 1.—*Critical Values for .10-Level One-Sided Test for Equality of Extreme-Value Scale Parameters*

| $n_1$ | $r$ | $p_1$ | $q_1$ | $n_2$ | $s$ | $p_2$ | $q_2$ | $Q_0$ | Power for | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | $b_1/b_2=.25$ | $b_1/b_2=.50$ |
| 6 | 4 | 1 | 4 | 5 | 2 | 1 | 2 | .591 | .492 | .267 |
| 6 | 4 | 1 | 4 | 5 | 3 | 1 | 3 | .411 | .601 | .306 |
| 6 | 4 | 1 | 4 | 5 | 4 | 1 | 4 | .330 | .663 | .328 |
| 6 | 4 | 1 | 4 | 5 | 5 | 1 | 5 | .271 | .709 | .345 |
| 6 | 4 | 1 | 4 | 6 | 2 | 1 | 2 | .601 | .489 | .265 |
| 6 | 4 | 1 | 4 | 6 | 3 | 1 | 3 | .422 | .595 | .303 |
| 6 | 4 | 1 | 4 | 6 | 4 | 1 | 4 | .345 | .653 | .324 |
| 6 | 4 | 1 | 4 | 6 | 5 | 1 | 5 | .295 | .694 | .339 |
| 6 | 4 | 1 | 4 | 6 | 6 | 1 | 6 | .251 | .728 | .352 |
| 6 | 5 | 1 | 5 | 2 | 2 | 1 | 2 | .661 | .561 | .303 |
| 6 | 5 | 1 | 5 | 3 | 2 | 1 | 2 | .737 | .535 | .291 |
| 6 | 5 | 1 | 5 | 3 | 3 | 1 | 3 | .489 | .685 | .353 |
| 6 | 5 | 1 | 5 | 4 | 2 | 1 | 2 | .773 | .525 | .286 |
| 6 | 5 | 1 | 5 | 4 | 3 | 1 | 3 | .538 | .661 | .341 |
| 6 | 5 | 1 | 5 | 4 | 4 | 1 | 4 | .419 | .746 | .380 |
| 6 | 5 | 1 | 5 | 5 | 2 | 1 | 2 | .793 | .519 | .283 |
| 6 | 5 | 1 | 5 | 5 | 3 | 1 | 3 | .562 | .649 | .335 |
| 6 | 5 | 1 | 5 | 5 | 4 | 1 | 4 | .457 | .725 | .368 |
| 6 | 5 | 1 | 5 | 5 | 5 | 1 | 5 | .379 | .782 | .396 |
| 6 | 5 | 1 | 5 | 6 | 2 | 1 | 2 | .807 | .516 | .282 |
| 6 | 5 | 1 | 5 | 6 | 3 | 1 | 3 | .577 | .642 | .332 |
| 6 | 5 | 1 | 5 | 6 | 4 | 1 | 4 | .477 | .714 | .363 |
| 6 | 5 | 1 | 5 | 6 | 5 | 1 | 5 | .410 | .764 | .387 |
| 6 | 5 | 1 | 5 | 6 | 6 | 1 | 6 | .352 | .806 | .408 |
| 6 | 6 | 1 | 6 | 2 | 2 | 1 | 2 | .858 | .584 | .319 |
| 6 | 6 | 1 | 6 | 3 | 2 | 1 | 2 | .954 | .556 | .304 |
| 6 | 6 | 1 | 6 | 3 | 3 | 1 | 3 | .644 | .723 | .382 |
| 6 | 6 | 1 | 6 | 4 | 2 | 1 | 2 | .999 | .544 | .298 |
| 6 | 6 | 1 | 6 | 4 | 3 | 1 | 3 | .706 | .696 | .366 |
| 6 | 6 | 1 | 6 | 4 | 4 | 1 | 4 | .555 | .792 | .417 |
| 6 | 6 | 1 | 6 | 5 | 2 | 1 | 2 | 1.025 | .538 | .295 |
| 6 | 6 | 1 | 6 | 5 | 3 | 1 | 3 | .737 | .683 | .359 |
| 6 | 6 | 1 | 6 | 5 | 4 | 1 | 4 | .604 | .769 | .402 |
| 6 | 6 | 1 | 6 | 5 | 5 | 1 | 5 | .504 | .832 | .440 |
| 6 | 6 | 1 | 6 | 6 | 2 | 1 | 2 | 1.042 | .534 | .293 |
| 6 | 6 | 1 | 6 | 6 | 3 | 1 | 3 | .756 | .675 | .355 |
| 6 | 6 | 1 | 6 | 6 | 4 | 1 | 4 | .629 | .756 | .395 |
| 6 | 6 | 1 | 6 | 6 | 5 | 1 | 5 | .545 | .813 | .427 |
| 6 | 6 | 1 | 6 | 6 | 6 | 1 | 6 | .470 | .858 | .457 |

* * *

# AN INVENTORY PROBLEM WITH OBSOLESCENCE*

William P. Pierskalla

*Southern Methodist University*
*Dallas, Texas*

**ABSTRACT**

A stochastic single product convex cost inventory problem is considered in which there is a probability, $\pi_j$, that the product will become obsolete in the future period $j$. In an interesting paper, Barankin and Denny essentially formulate the model, but do not describe some of its interesting and relevant ramifications. This paper is written not only to bring out some of these ramifications, but also to describe some computational results using this model. The computational results show that if obsolescence is a distinct possibility in the near future, it is quite important that the probabilities of obsolescence be incorporated into the model before computing the optimal policies.

## 1. INTRODUCTION

In certain inventory situations it is the case that technological change or changes in the techniques of production may make a product obsolete almost overnight. In some of these cases it may be possible to state a probability mass function (pmf) which gives the probability $\pi_j$ that obsolescence will occur in some period $j$ in the future. $\pi_j$ is the pmf of the random variable $N$ which describes the length of the horizon. Thus the primary inventory problem to be considered here is the stochastic single-product convex-cost $N$ period problem, where the number of periods $N$ is a random variable. In an interesting paper, Barankin and Denny [3] essentially formulate the model, but do not describe some of its interesting and relevant ramifications.

This paper is written not only to bring out some of these ramifications, but also to describe some computational results using this model versus the well known $n$ period model of Arrow, Harris, and Marschak [1] and Bellman, Glicksberg, and Gross [6] and the infinite period model so capably presented in Arrow, Karlin, and Scarf [2].

The author's original interest in this model was stimulated by a paper of Professor Amnon Rapoport [10]. Rapoport describes an experiment he performed with a group of students in an attempt to determine their abilities to think dynamically. The students who had not had an inventory course were confronted with a sequence of six stochastic, single-product, convex-cost, inventory problems and at each stage they had to decide how much inventory to order. He compared their performance with the infinite stage model and found that they did not perform particularly well. Unfortunately he did not use the correct model. These students knew that no problem would last an infinite number of periods; furthermore the actual number of periods, to them, was a random variable independent of the demands in each period. Hence, in this case also, the appropriate inventory model is the obsolescence situation soon to be described in detail.

*Research supported by the National Science Foundation, Grant No. GK–1333.

In the next section the model formulation is given. In the subsequent section some characteristics of the model are presented and some particular obsolescence distributions are discussed. Then in the final section some computational results are given, which in themselves are worthy of note due to the manner in which the critical numbers behave for various types of obsolescence distributions.

## 2. MODEL FORMULATION

The following definitions, assumptions, and conventions will be used:

(a) In discussing the finite period problem the periods will be numbered backwards. The numbers $M$ and $T$ are integers representing the maximum number of periods that the problem will last and the first period in which the probability of obsolescence $\pi_T$ first becomes positive, respectively $(T \leq M)$.

(b) $h(\cdot)$ are $p(\cdot)$ are continuously differentiable convex increasing holding and shortage cost functions, respectively, and will be charged at the end of the period after the period's demand has occurred, but prior to delivery of stock for the next period. It is assumed $h(0)=p(0)=0$.

(c) There is no backlogging of excess demand.

(d) Delivery is immediate.

(e) $c$ is the marginal cost of one unit of stock purchased, $r$ is the marginal revenue of one unit of stock sold $(r > c)$, and $\alpha$ is the discount factor $0 \leq \alpha \leq 1$.

(f) $\pi_j$ is the probability of obsolescence in period $j$ after ordering and after the occurrence of demand $(j=1, \ldots, M)$. Let $w$ denote the period in which obsolescence occurs and let $p_r$ be the conditional probability that the problem terminates in period $r \leq t \leq m$ given that the problem has not terminated in periods $M, M-1, \ldots, t+1$. Then

$$P_r = P\{w=r \mid w \leq t\}$$

$$= \frac{P\{w=r \text{ and } w<t\}}{P\{w \leq t\}} = \frac{\pi_r}{\sum_{k=1}^{t} \pi_k} \quad ; \quad r=1, \ldots, t.$$

Of course $1-p_r$ is the conditional probability that the problem will not terminate in period $r$, given it has survived to period $t$.

(g) $\phi(\xi)$ is the probability density function $(pdf)$ of the nonnegative demand random variable D. The demands are assumed to be independent and identically distributed among the periods.

(h) $$L(y) = \int_0^y h(y-\xi)\phi(\xi)d\xi - r\int_0^y \xi\phi(\xi)d\xi - ry\int_y^\infty \phi(\xi)d\xi + \int_y^\infty p(\xi-y)\phi(\xi)d\xi. \tag{1}$$

(i) $\bar{f}_j(x)$ is the minimum discounted conditional expected cost from period $j$ onward given the initial stock on hand in period $j$ is $x$ and given that obsolescence has not occurred in prior periods.

(j) $$\bar{G}_j(y) = cy + L(y) + \alpha(1-p_j)\left[\bar{f}_{j-1}(0)\int_y^\infty \phi(\xi)d\xi + \int_0^y \bar{f}_{j-1}(y-\xi)\phi(\xi)d\xi\right]. \tag{2}$$

(k) $$\bar{f}_j(x) = \min_{y \geq x}\{\bar{G}_j(y) - cx\}.$$

(1) It is assumed that $h'(0)+p'(0)+r \geq \alpha c(1-p_n)$ for all $n=1,2, \ldots$. This condition holds if $h'(0)+p'(0)+r \geq \alpha c$ as required and justified in [2], [6], and others. Furthermore, it is assumed that $E[p'(D)] > c-r$, where $p'(x)=dp(x)/dx$. This latter condition ensures that the one period problem has a positive finite solution $y_1^*$.

The inventory problem is to find $y \geq x$ to minimize $\bar{G}_j(y)-cx$. This problem as well as the Arrow, Harris, and Marschak [1] problem has a single critical number independent of the initial stock $x$.

## . CHARACTERISTICS OF THE MODEL

Some of the characteristics of the model are independent of the type of distribution chosen for $\pi_j\}$. In these cases the same conclusions hold as in the Bellman, Glicksberg, and Gross [6] and Arrow, Karlin, and Scarf [2] models and for the same reasons. Other results, however, are dependent on the nature of the $\{\pi_j\}$. In the former cases the results will be given without proofs (proofs for the various parts are available in [8], [7], [6], [1], [2]) and in the latter cases proofs will be supplied in the appendix.

*THEOREM 1:* $\bar{f}_n(x)$ is a convex function and the optimal policy in each period for the $n$-period problem is a single critical number, $y^*(p_n, p_{n-1}, \ldots, p_1) \equiv y^*(P_n)$, where $y^*(P_n)$ is defined as the smallest number satisfying $\bar{G}_n'(y)=0$.

*LEMMA 1:* $\bar{f}_n'(x)=\dfrac{d\bar{f}_n(x)}{dx} \geq -c$ for all $x$, and all $n$.

Let $p=\lim_{n\to\infty} p_n$, $\quad (0 \leq p \leq 1)$, $\quad$ and let $P_\infty=(p_1, p_2, p_3, \ldots)$.

Define $y^*(P_\infty)$ as the smallest number $y$ satisfying

(3) $$c+L'(y)+\alpha(1-p)\int_0^y \bar{f}'(y-\xi)\phi(\xi)d=0$$

i.e., $c+L'(y)-\alpha c(1-p)\phi(y)=0)$, where $\bar{f}(x)$ is the minimum discounted expected costs for an infinite horizon given an initial stock of $x$ units and given that obsolescence has not occurred. When discussing the infinite period case the periods will be numbered forward.

It will be assumed that $L'(y)=0$ has a solution and the smallest $y$ satisfying $L'(y)=0$ will be denoted by $\bar{y}$. Define

(4) $$\bar{G}(y)=c\cdot y+L(y)+\alpha(1-p)[\bar{f}(0)\int_y^\infty \phi(\xi)d\xi+\int_0^y \bar{f}(y-\xi)\phi(\xi)d\xi].$$

*THEOREM 2:* If $p>0$ or if $\alpha<1$, or both, then

(a) $\lim_{n\to\infty} \bar{f}_n=\bar{f}(x)$: $\lim_{n\to\infty} G_n(x)=\bar{G}(x)$,

(5) (b) $\bar{f}(x)=\min_{y\geq x}\{\bar{G}(y)-cx\}$

(c) $\bar{f}(x)$ is a convex function,

(d) $y^*(P_\infty)$ is an optimal solution to (5), and

(e) the sequence $\{y^*(P_n)\}$ contains convergence subsequences and every limit point of $\{y^*(P_n)\}$ satisfies (3). Furthermore if (3) has a unique solution, then $\{y^*(P_n)\}$ converges to $y^*(P_\infty)$. Equation (3) has a unique solution if $L(y)$ is strictly convex.

The following results hold under certain assumptions on the probabilities

$$[\pi_j]_{j=1}^{M}$$

It will be assumed that the random variable $N$ has an increasing failure rate (IFR) distribution. The definition of IFR is as follows: A discrete random variable has an increasing failure rate distribution provided the function

$$p(j)=\frac{\pi_j}{\sum_{i=j+1}^{\infty}\pi_j}$$

is increasing in $j$. In an interesting paper, Barlow, Marshall, and Proschan [4] describe many properties of IFR distributions. Among these properties is the following: "If $N$ is a time variable and time is reversed, then the random variable $N$ has an increasing failure rate if, and only if, $-N$ has the decreasing ratio

$$\pi_j/\sum_{i=1}^{j}\pi_i."$$

For this inventory model then $N$ has an IFR distribution if, and only if, $p_j$ is decreasing, since time has been reversed in the $n$-period model.

*THEOREM 3:* If $1-p_j$ is a nondecreasing function of $j$ then

(a) $y^*(P_1) \leq y^*(P_2) \leq \ldots \leq y^*(P_n) \leq y^*(P_\infty) \leq \bar{y}$.

(b) $\lim_{n\to\infty} y^*(P_n)=y^*(P_\infty)\leq \bar{y}$.

Some examples of IFR distributions which might prove useful in this obsolescence context are:

(a) the equally likely distribution

$$\pi_j=\begin{cases}\frac{1}{T} \text{ for } j=1, \ldots, T\\ 0 \text{ otherwise.}\end{cases}$$

(b) the truncated Poisson distribution

$$\pi_j=\begin{cases}\left(\frac{\lambda^{j-1}}{(j-1)!}\right)\cdot\left(\sum_{k=1}^{T}\frac{\lambda^{k-1}}{(k-1)!}\right)^{-1} & \text{for } j=1, \ldots, T\\ & \lambda>0\\ 0 & \text{otherwise.}\end{cases}$$

(c) the discrete triangular distribution

$$\pi_j=\begin{cases} j\left(\sum_{k=1}^{T} k\right)^{-1} & \text{for } j=1, \ldots, T\\ 0 & \text{otherwise.}\end{cases}$$

(d) the truncated geometric distribution

$$\pi_j=\begin{cases}\gamma^{j-1}\left(\sum_{k=1}^{T}\gamma^{k-1}\right)^{-1} & \text{for } j=1, \ldots, T\\ & 0<\gamma<1\\ 0 & \text{otherwise.}\end{cases}$$

Actually pmf (*d*) would perhaps be less reasonable since it weights the very last periods most heavily; however, there could be obsolescence situations in which the last periods should receive the greatest weights. It will be seen later on that because of this property pmf (*d*) behaves more like the *n*-period model (*n* known) than any of the distributions (*a*), (*b*), or (*c*).

There are many other IFR distributions besides the foregoing. The book by Barlow and Proschan [5] gives a listing and discussion of them.

Of these four distributions perhaps the most useful is the truncated Poisson since it allows one to choose the parameters $\lambda$ and $T$ in such a way as to locate the mean and range of the distribution as desired. In this case,

$$E[N]=\frac{\lambda \sum_{j=1}^{T-1} \frac{\lambda^{j-1}}{(j-1)!}}{\sum_{j=1}^{T} \frac{\lambda^{j-1}}{(j-1)!}}$$

and

$$E[N^2]=\frac{\lambda \sum_{j=1}^{T-1} \frac{\lambda^{j-1}}{(j-1)!} + \lambda^2 \sum_{j=1}^{T-2} \frac{\lambda^{j-1}}{(j-1)!}}{\sum_{j=1}^{T} \frac{\lambda^{j-1}}{(j-1)!}}.$$

Distributions (*a*) and (*c*) because of their single parameter $T$ do not have the same flexibility. Of course many other multi-parameter distributions for the nonnegative random variable $N$ could be used (see [5]).

For these four distributions the conditional probabilities of obsolescence, $p_j$'s, *given the process has j periods to go* are

(a) $p_j=\frac{1}{j}$ $\qquad j=1, \ldots, T$

(b) $p_j=\left(\frac{\lambda^{j-1}}{(j-1)!}\right)\left(\sum_{k=1}^{j} \frac{\lambda^{k-1}}{(k-1)!}\right)^{-1}$ $\qquad j=1, \ldots, T$, $\lambda>0$

(c) $p_j=\frac{2}{j+1}$ $\qquad j=1, \ldots, T$

(d) $p_j=\frac{\gamma^{j-1}}{\sum_{k=1}^{j} \gamma^{k-1}}$ $\qquad j=1, \ldots, T$, $0<\gamma<1$.

It should be noted that as $T\to+\infty$ we are numbering foreward and the distributions (b) and (d) become the Poisson and geometric distributions respectively. Furthermore for

(*b*) $p=e^{-\lambda}>0$ $\qquad$ for $0<\lambda<\infty$,
(d) $p=1-\gamma>0$ $\qquad$ for $0<\gamma<1$,

and (a) and (c) $p=0$.

The foregoing results, Lemma 1 and Theorems 1, 2, and 3, hold under various modifications of the assumptions of the model. If assumption (c) is removed and complete backlogging of excess demand is allowed, then lags in delivery are permissible. It is also possible to obtain some bounds on the critical numbers $y^*(P_n)$ which indicate the rate of convergence of the $y^*(P_n)$ to $y^*(P_\infty)$ when IFR distributions are used for $\{\pi_j\}$. Before stating the bounds, it is useful to give the following Lemma.

*LEMMA 2:* If backlogging is allowed and if $1-p_j$ is nondecreasing in $j$ then for any $a \geq 0$ and all $x$ such that $x \leq x+a \leq y^*(P_\infty)$, $0 \leq f_n(x+a)-f_n(x)+ac \leq \mu_n(a)$, $n=0, 1, \ldots$ where

$$\mu_0(a)=ac$$

$$\mu_n(a)=\alpha[ac(1-p)+(1-p_n)(\mu_{n-1}(a)-ac)].$$

The lemma states that for any two points in the domain lying below the critical number for the infinite problem $y^*(P^\infty)$ call these points $x_1$ and $x_2$, where $x_2-x_1=a \geq 0$, the difference in the functional, values $f(x_2)-f(x_1)$ is bounded by

$$-c(x_2-x_1) \leq f_n(x_2)-f_n(x_1) \leq \mu_n(x_2-x_1)-c(x_2-x_1).$$

It is not difficult to show by induction that $\mu_n(a)$ can be given by the following formula. First define $1-p_j \equiv q_j$ and $1-p=q$ then

$$\mu_n(a)=ac\left[\alpha(q-q_n)+\sum_{j=2}^{n}\alpha^j(q-q_{n-j+1})\prod_{k=1}^{j-1}q_{n-k+1}+\alpha^n\prod_{k=1}^{n}q_{n-k+1}\right]$$

for all $n=1, 2, 3, \ldots, \mu_0(a)=ac$.

For Theorem 3, if $1-p_j$ is nondecreasing then $y^*(P_n) \leq y^*(P_\infty)$ and it is perfectly acceptable to let $y^*(P_\infty)-y^*(P_n)=a$ in Lemma 2 above.

*LEMMA 3:* If backlogging is allowed, $L(y)$ is a convex function $\epsilon C^2$, $L''(y)>0$ for $y\epsilon[0, \bar{y}]$, and $1-p_j$ is nondecreasing in $j$, then

$$0 \leq y^*(P_\infty)-y^*(P_n) \leq \frac{c}{L''(\hat{y})}\left[\alpha(q-q_n)+\sum_{j=2}^{n}\alpha^j(q-q_{n-j+1})\prod_{k=1}^{j-1}q_{n-k+1}+\alpha^n\prod_{k=1}^{n}q_{n-k+1}\right],$$

where $\hat{y}$ is the point in the open interval $(y^*(P_n), y^*(P_\infty))$ given by the second order Taylor expansion

$$L(y^*(P_\infty))=L(y^*(P_n))+aL'(y^*(P_n))+\frac{a^2}{2}L''(\hat{y}).$$

Other modifications to the assumptions are possible. For example, a fixed setup cost $K \cdot \delta(y-x)$ may be added to the production cost

$$\left(\delta(y-x)=\begin{cases}1 \text{ if } y>x\\ 0 \text{ if } y=x\end{cases}\right).$$

In this case the optimal policies will be $(s, S)$ and the cost functions $K$-convex; however, the preceding results will not be valid.

## 4. COMPUTATIONAL RESULTS

The results shown in this section are quite revealing. In general they show that if obsolescence is a distinct possibility in the near future, it is important that the probabilities of obsolescence be incorporated into the model before computing the optimal $M$ period policies.

The computations were obtained on the Case Western Reserve University Univac 1107. In the dynamic programming algorithm the search for the optimal policy $y^*(P_j)$ in each stage was confined to integers and lattice search was used [11]. For this reason the actual $y^*(P_j)$'s so obtained may vary slightly from the true $y^*(P_j)$'s.

Two examples were run. The following parameters were chosen.

| | $c$ | $h$ | $p$ | $r$ | $M=T$ | Demand Erlang | Expected demand | Standard deviation of demand |
|---|---|---|---|---|---|---|---|---|
| Example 1... | \$17.80 | \$2.10 | \$46.00 | \$22.50 | 15 | $r=11.0$ $\beta=0.044$ | 250 | 75 |
| Example 2... | \$17.80 | \$2.10 | \$46.00 | \$22.50 | 25 | $r=4.0$ $\beta=1/3$ | 12 | 6 |

The costs above coincide with the costs for Rapoport's [10] Example Number 4. Other costs (not given here) were run which yield the same type of results. The Erlang distribution was chosen for the demand distribution since it is a two parameter distribution:

$$\phi(\xi)=\frac{\beta}{(r-1)!}(\beta\xi)^{r-1}e^{-\beta\xi} \qquad \text{for } \xi \geq 0$$

$$=0 \qquad \text{otherwise.}$$

The critical numbers for each example were computed using the $M$ period problem without obsolescence and using the four obsolescence distributions (a)–(d) above. In distribution (b), the truncated Poisson distribution, $\lambda$ was chosen $\lambda=10$; in distribution (d), the truncated geometric distribution, $\gamma$ was chosen $\gamma=1/2$.

The critical numbers are portrayed on the following two graphs (Figures 1 and 2).

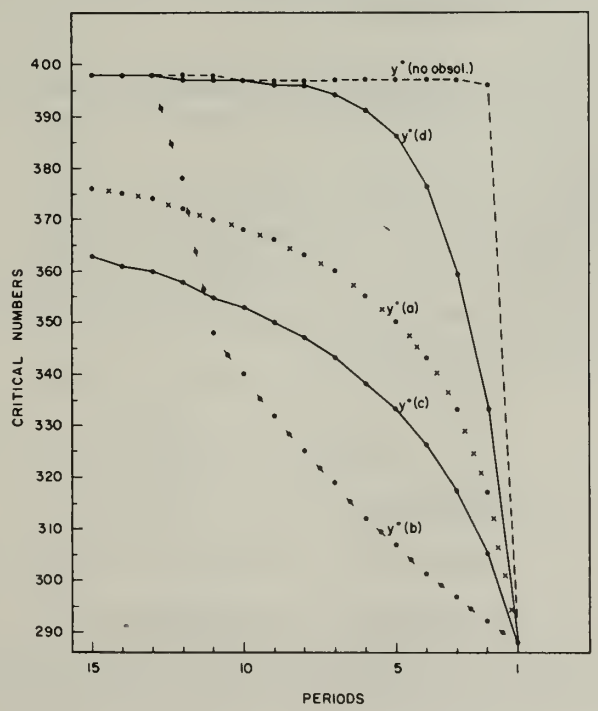



Figure 1. Example 1: Critical Numbers— Demand is Gamma (11.0, 0.044)

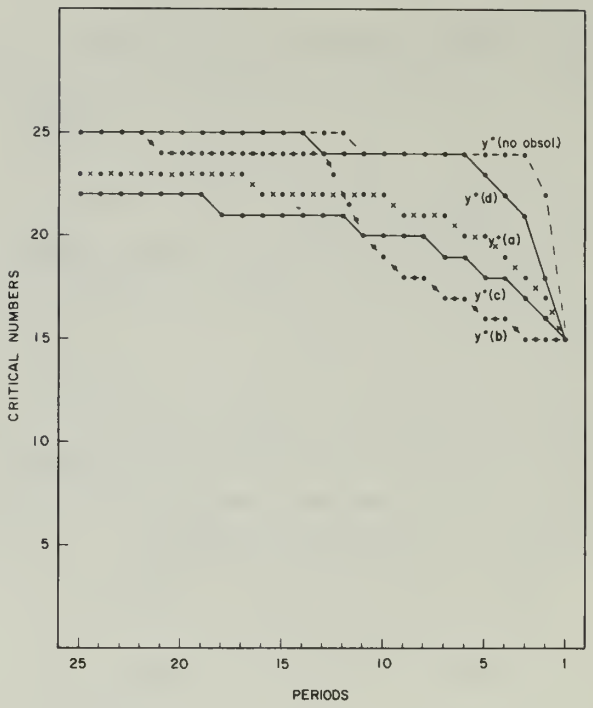



Figure 2. Example 2: Critical Numbers—Demand is Gamma (4, 1/3)

It is apparent in both examples that the critical numbers without obsolescence rapidly approach the steady state number $y^*$. Furthermore the critical numbers for the truncated geometric are close to the critical numbers for the no obsolescence case except near the end of the process. As mentioned earlier, this phenomenon is caused by the heavy weighting attached to the ending periods by the truncated geometric distribution.

On the other hand, the other three obsolescence distributions (a), (b), and (c) differ markedly from the no obsolescence results and indeed over the range of $M$ studied here, distributions (a) and (c) have not achieved the steady state.

The conclusion reached by these examples is that it is important to consider obsolescence probabilities in computing the critical numbers for an inventory problem. This conclusion is particularly applicable to spare parts stocking problems where the threat of obsolescence is high.

As a final point, an example is given below where an IFR distribution for $\pi$ is not used and the critical numbers are not ordered as in Theorem 3.

| | $c$ | $h$ | $p$ | $r$ | $M=\mathrm{T}$ | Demand Erlang | $\pi_1$ | $\pi_2$ | $\pi_3$ |
|---|---|---|---|---|---|---|---|---|---|
| Example 3 | \$17.80 | \$2.10 | \$46.00 | \$22.50 | 3 | $\gamma=4.0$ $\lambda=1/3$ | 0.019 | 0.001 | 0.980 |

By using these parameters we find the optimal three period policies to be $y^*(P_3)=15$, $y^*(P_2)=21$, and $y^*(P_1)=15$, and we do not have the ordering given in Theorem 3; however, the lack of ordering among the $y^*(P_j)$'s in no way detracts from the importance of considering the obsolescence probabilities when computing the critical numbers.

## Appendix

*PROOF OF THEOREM 3:*

(a) The theorem will be proved by induction on $n$. We first show $y^*(P_1) \leq y^*(P_2)$. Recall that $y^*(P_1)$ and $y^*(P_2)$ are the smallest $y$'s satisfying $\bar{G}_1'(y) = L'(y) + c = 0$ and $\bar{G}_2'(y) = c + L'(y) + \alpha(1-p_2) \int_0^y \bar{f}_1'(y-\xi)\phi(\xi)d\xi = 0$. Furthermore, $L'(0) = -r - E[p'(D)] < -c = L'(y^*(P_1))$ and since $L(\cdot)$ is convex $y^*(P_1) > 0$. Consider any $y$ in the right-closed interval $(0, y^*(P_1)]$.

$$\begin{aligned}\bar{G}_2'(y) - \bar{G}_1'(y) &= \alpha(1-p_2) \int_0^y \bar{f}_1'(y-\xi)\phi(\xi)d\xi \\ &= \alpha(1-p_2) \int_0^y (-c)\phi(\xi)d\xi \\ &= -\alpha c(1-p_2)\phi(y) \leq 0.\end{aligned}$$

Hence $\bar{G}_2'(y) \leq \bar{G}'(y)$ for all $y \in (0, y^*(P_1)]$ and $\bar{G}_2'(y^*(P_1)) \leq \bar{G}_1'(y^*(P_1)) = 0$. Since $\bar{G}_2$ is convex, then $y^*(P_2) \geq y^*(P_1)$. In this case the proof is independent of the behavior of $p_2$.

We will now show $y^*(P_n) \geq y^*(P_{n-1})$. Again note $y^*(P_n)$ is obtained as the smallest $y$ satisfying $\bar{G}_n'(y) = 0$. $y^*(P_n)$ exists since $\bar{G}_n'(0) = L'(0) < 0$ and $\bar{G}_n(+\infty) = +\infty$ and $\bar{G}_n$ is convex. Similarly $y^*(P_{n-1})$ is obtained as the smallest $y$ satisfying $\bar{G}_{n-1}'(y) = 0$ and it also exists. Now

$$\bar{G}_n'(y) - \bar{G}_{n-1}'(y) = \alpha[(1-p_n) \int_0^y \bar{f}_{n-1}'(y-\xi)\phi(\xi)d\xi - (1-p_{n-1}) \int_0^y \bar{f}_{n-2}'(y-\xi)\phi(\xi)d\xi].$$

Furthermore, for all $y$ such that $y \leq y^*(P_{n-1})$

$$\bar{f}_{n-1}'(y) = -c$$

and $\bar{f}_{n-1}'(y-\xi) = -c$ for all $\xi \geq 0$

From lemma $1 \bar{f}_{n-2}'(x) \geq -c$ for all $x$. Thus for all $y$ such that $y \leq y^*(P_{n-1})$

$$\begin{aligned}&\bar{G}_n'(y) - \bar{G}_{n-1}'(y) \\ &\quad = \alpha[(1-p_n)(-c)\Phi(y) - (1-p_{n-1}) \int_0^y \bar{f}_{n-2}'(y-\xi)\phi(\xi)d\xi] \\ &\quad \leq \alpha[(1-p_n)(-c)\Phi(y) + (1-p_{n-1}) \int_0^y c\phi(\xi)d\xi] \\ &\quad = -\alpha c\Phi(y)[(1-p_n) - (1-p_{n-1})] \\ &\quad \leq 0 \text{ since } 1-p_j \text{ is nondecreasing in } j.\end{aligned}$$

Then $\bar{G}_n'(y) \leq \bar{G}_{n-1}'(y)$ for all $y \leq y^*(P_{n-1})$; in particular,

$$\bar{G}_n'(y^*(P_{n-1})) \leq \bar{G}_{n-1}'(y^*(P_{n-1})) = 0$$

and since $\bar{G}_j(y)$ is convex for all $j$, $y^*(P_n) \geq y^*(P_{n-1})$. We now show

$$y^*(P_n) \leq y^*(P_\infty) \qquad \text{for all } n.$$

The proof is the same as given above for the $n$ period case with $\bar{G}'$ replacing $\bar{G}_n'$ and $\bar{G}_n'$ replacing $\bar{G}_{n-1}'$.

Finally, we show $y^*(P_\infty) \leq \bar{y}$.

$$L'(y^*(P_\infty)) = -c - \alpha(1-p)\int_0^{y^*(P_\infty)} \bar{f}'(y^*(P_\infty)-\xi)\phi(\xi)d\xi$$

$$= -c + \alpha c(1-p)\Phi(y^*(P_\infty))$$

$$\leq -c + \alpha c(1-p) < 0 = L'(\bar{y}).$$

Since $L(\cdot)$ is convex, $y^*(P_\infty) \leq \bar{y}$.

(b) Since the $\{y^*(P_n)\}$ is monotonic by part (a) above and bounded from above by $y^*(P_\infty)$, it has a limit point $y^*$. We must show $y^* = y^*(P_\infty)$. (Since we already know $y^* \leq y^*(P_\infty)$, we only must show $y^* \not< y^*(P_\infty)$.) Assume $y^* < y^*(P_\infty)$. Since $y^*(P_\infty)$ is the smallest $y$ satisfying $\bar{G}'(y) = 0$, then $\bar{G}'(y^*) < \bar{G}'(y^*(P_\infty)) = 0$; however, we know $\bar{G}_n'(y^*(P_n)) = 0$ for all $n$ and $\lim_{n\to\infty} \bar{G}_n'(x) = \bar{G}'(x)$. Thus by the uniform continuity of $\bar{G}_n'$ on $[0, \bar{y}]$, we have $\lim_{n\to\infty} \bar{G}_n'(y^*(P_n)) = \bar{G}'(y^*) = 0$ which contradicts $\bar{G}'(y^*) < 0$. Hence $y^* = y^*(P_\infty)$. Q.E.D.

*PROOF OF LEMMA 2:*

Since $\bar{f}_n(\cdot)$ is convex and $\bar{f}_n(\cdot) \geq -c$, then for $a \geq 0$, $\bar{f}_n(x+a) - \bar{f}_n(x) + ac \leq \bar{f}_n(x+a) - \bar{f}_n(x) - a\bar{f}'(x) \leq 0$ for all $n$. We must show that the second inequality holds. It is clearly true for $n = 0$. Hence assume it is true for $n-1$. We will consider three cases. ($\mu_n(a)$ will be abbreviated by $\mu_n$.)

*CASE 1:* $y_n^* \leq x \leq x+a \leq y^*$

$$\bar{f}_n(x+a) - \bar{f}_n(x) + ac$$

$$= L(x+a) - L(x) + ac + \alpha q_n \int_0^\infty [\bar{f}_{n-1}(x+a-\xi) - \bar{f}_{n-1}(x-\xi)]\phi(\xi)d$$

$$\leq L(x+a) - L(x) + ac + \alpha q_n \int_0^\infty [\mu_{n-1} - ac]\phi(\xi)d\xi$$

$$= L(x+a) - L(x) + ac + \alpha q_n[\mu_{n-1} - ac]$$

$$\leq aL'(y^*) + ac + \alpha q_n[\mu_{n-1} - ac] \text{ since } L(\cdot) \text{ is convex}$$

$$= ac[\alpha q - 1] + ac + \alpha q_n[\mu_{n-1} - ac]$$

$$= \alpha[acq + q_n(\mu_{n-1} - ac)] = \mu_n \text{ as required.}$$

*CASE 2:* $x \leq y_n^* \leq x+a \leq y^*$

$$\bar{f}_n(x+a) - \bar{f}_n(x) + ac$$

$$= L(x+a) + \alpha q_n \int_0^\infty \bar{f}_{n-1}(x+a-\xi)\phi(\xi)d\xi$$

$$- c(y_n^* - x) - L(y_n^*) - \alpha q_n \int_0^\infty \bar{f}_{n-1}(y_n^* - \xi)\phi(\xi)d\xi + ac$$

$$= L(x+a) - L(y_n^*) + ac - c(y_n^* - x)$$

$$+ \alpha q_n \int_0^\infty [\bar{f}_{n-1}(x+a-y_n^*+y_n^*-\xi) - \bar{f}_{n-1}(y_n^*-\xi)]\phi(\xi)d\xi$$

$$\leq L(x+a) - L(y_n^*) + ac - c(y_n^* - x)$$

$$+ \alpha q_n \int_0^\infty [\mu_{n-1}(x+a-y_n^*) - (x+a-y_n^*)c]\phi(\xi)d\xi$$

$$= L(x+a) - L(y_n^*) + ac - c(y_n^* - x) + \alpha q_n[\mu_{n-1}(x+a-y_n^*) - (x+a-y_n^*)c]$$

$$\leqslant aL'(y^*) + c(x+a-y_n^*) + \alpha q_n[\mu_{n-1}(x+a-y_n^*) - (x+a-y_n^*)c]$$

$$= ac[\alpha q - 1] + c(x+a-y_n^*) + \alpha q_n[\mu_{n-1}(x+a-y_n^*) - (x+a-y_n^*)c]$$

$$= ac[\alpha q - 1] + c(x+a-y_n^*)$$

$$+ \alpha q_n[(x+a-y_n^*)c\alpha(q-q_{n-1})$$

$$+ (x+a-y_n^*)c\sum_{j=2}^{n-1}\alpha^j(q-q_{n-j})\prod_{k=1}^{j-1}q_{n-k}$$

$$+ (x+a-y_n^*)c\alpha^{n-1}\prod_{j=1}^{n-1}q_{n-j} - (x+a-y_n^*)c]$$

Since $q \geqslant q_t \geqslant 0$ and $x - y_n^* \leqslant 0$, then $x+a-y_n^* \leqslant a$

$$\leqslant ac[\alpha q - 1] + ac + \alpha q_n\Big[ac\alpha(q-q_{n-1})$$

$$+ ac\sum_{j=2}^{n-1}\alpha^j(q-q_{n-j})\prod_{k=1}^{j-1}q_{n-k}$$

$$+ ac\alpha^{n-1}\prod_{k=1}^{n-1}q_{n-k} - ac\Big]$$

$$+ c(x-y_n^*) - \alpha q_n c(x-y_n^*).$$

(Now, $c(x-y^*) - \alpha q_n c(x-y^*) \leqslant 0$.)

$$\leqslant \alpha[acq + q_n\Big\{ac\alpha(q-q_{n-1}) + ac\sum_{j=2}^{n-1}\alpha^j(q-q_{n-j})\prod_{k=1}^{j-1}q_{n-k}$$

$$+ ac\alpha^{n-1}\prod_{k=1}^{n-1}q_{n-k} - ac\Big\}$$

$$= \alpha[acq + q_n(\mu_{n-1} - ac)] = \mu_n \text{ as required.}$$

*CASE 3:* $x \leqslant x+a \leqslant y_n^* \leqslant y^*$. On this interval $\bar{f}(y) = \bar{f}_n'(x+a) = \bar{f}_n'(x) = -c$ for $y \in [x, x+a]$. Thus $\bar{f}_n(y)$ is linear on $[x, x+a]$, *and we have*

$$\frac{\bar{f}_n(x+a) - \bar{f}_n(x)}{x} = -c,$$

i.e.,

$$\bar{f}_n(x+a) - \bar{f}_n(x) + ac = 0 \leqslant \mu_n. \qquad \text{Q.E.D.}$$

*PROOF OF LEMMA 3:*

Define $y^* \equiv y^*(P_\infty)$, $y_n^* \equiv y^*(P_n)$,

$$\delta_n = \alpha(q-q_n) + \sum_{j=2}^{n}\alpha^j(q-q_{n-j+1})\prod_{k=1}^{j-1}q_{n-k+1} + \alpha^n\prod_{j=1}^{n}q_{n-j+1},$$

$a = y^* - y_n^*$. Thus since $\bar{f}_n$ and $L$ are convex:

$$\mu_n(a) \geq \bar{f}_n(y^*) - \bar{f}_n(y_n^*) + ac$$

$$= L(y^*) + \alpha q_n \int_0^\infty \bar{f}_{n-1}(y^* - \xi)\phi(\xi)\,d\xi$$

$$- L(y_n^*) - \alpha q_n \int_0^\infty \bar{f}_{n-1}(y_n^* - \xi)\phi(\xi)\,d\xi + ac$$

$$\geq L(y^*) - L(y_n^*) - aL'(y_n^*)$$

$$+ \alpha q_n \int_0^\infty [\bar{f}_{n-1}(y_n^* - \xi) - \bar{f}_{n-1}(y_n^* - \xi) - a\bar{f}'_{n-1}(y_n^* - \xi)]\phi(\xi)\,d\xi$$

$$\geq L(y^*) - L(y_n^*) - aL'(y_n^*)$$

$$= \frac{a^2}{2} L''(\hat{y}) \text{ where } \hat{y} \in [y_n^*, y^*].$$

Hence

$$\frac{a^2}{2} L''(\hat{y}) \leq \mu_n(a) = ac\delta_n$$

or

$$\frac{a}{2} L''(\hat{y}) \leq c\delta_n \qquad \text{implies}$$

$$0 \leq a = y^* - y_n^* \leq \frac{2c\delta_n}{L''(\hat{y})}. \qquad \text{Q.E.D.}$$

## REFERENCES


[1] Arrow, K., T. Harris, and J. Marschak, "Optimal Inventory Policy," *Econometrica*, *19*, 250–272 (1951).

[2] Arrow, K. J., S. Karlin, and H. Scarf, *Studies in the Mathematical Theory of Inventory and Production* (Stanford University Press, Stanford, California, 1958).

[3] Barankin, E. W. and J. Denny, "Examination of an Inventory Model Incorporating Probabilities of Obsolescence," *Logistics Review and Military Logistics Journal*, *1*, 11–25 (1965).

[4] Barlow, R. E., A. W. Marshall, and F. Proschan, "Properties of Probability Distributions with Monotone Hazard Rate," *Annals of Math. Statistics*, *34*, 375–389 (1963).

[5] Barlow, R. E. and F. Proschan, *Mathematical Theory of Reliability* (John Wiley and Sons, New York, 1965).

[6] Bellman, R., I. Glicksberg, and O. Gross, "On the Optimal Inventory Equation," *Management Science*, *2*, 83–104 (1955).

[7] Iglehart, D. L., "Dynamic Programming and Stationary Analysis of Inventory Problems," Chapter 1 in *Multistage Inventory Models and Techniques* (H. Scarf, D. Gilford, and M. Shelly, Editors) (Stanford University Press, Stanford, California, 1963).

[8] Iglehart, D. L., "Optimality of (s, S) Policies in the Infinite Horizon Dynamic Inventory Problem," *Management Science*, *9*, 259–267 (1963).

[9] Pierskalla, W. P., "Analysis of a Multistage Inventory Task—Comments on a Paper of Amnon Rapoport," Technical Memorandum No. 98, February, 1968, Operations Research Department, Case Western Reserve University, Cleveland, Ohio.

[10] Rapoport, Amnon, "Variables Affecting Decisions in a Multistage Inventory Task," *Behavioral Science*, *12*, 194–204 (1967).

[11] Wilde, D. J., *Optimum Seeking Methods* (Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1964).


* * *

# WAR RESERVE SPARES KITS SUPPLEMENTED BY NORMAL OPERATING ASSETS

*Robin B. S. Brooks* and John Y. Lu*

*The RAND Corporation*
*Santa Monica, California*

**ABSTRACT**

In peacetime, base stock levels of spares are determined on the assumption of normal resupply from the depot. In the event of war, however, a unit must be prepared to operate from stock on hand for a period of time without being resupplied from the depot. This paper describes a mathematical model for determining such war reserve spares (WRS) requirements. Specifically, the model solves the following kind of optimization problem: find the least-cost WRS kits that will keep the probability of a stockout after $K$ cannibalizations less than or equal to some target objective $\alpha$. The user of the model specifies the number of allowable cannibalizations, and the level of protection that the kit is supposed to provide.

One interesting feature of this model is that in the probability computation it takes into account the possiblility of utilizing normal base operating assets. Results of a sensitivity analysis indicate that if peacetime levels were explicitly taken into account when designing a WRS kit, a cost saving of nearly 40 percent could be effected without degrading base supply performance in wartime.

## I. INTRODUCTION

This paper describes a mathematical model that was used to assist the Air National Guard in designing war reserve spares (WRS) kits. It extends an earlier effort for the Aerospace Defense Command (ADC) in which RAND developed a model to compute WRS kits that ADC would use at their dispersal sites. In the summer of 1966, ADC, together with the Air National Guard, asked RAND to help them construct WRS kits to be used by Air National Guard units assigned to an air defense mission. The Air National Guard operates under a fight-in-place concept, rather than a dispersal concept that ADC uses. This difference should be reflected in the design of WRS kits, because the Air National Guard can supplement their war reserve spares in the event of war with assets that they use for normal day-to-day operations. It seemed, therefore, that WRS kits for the Air National Guard could be designed at considerably smaller cost than kits used by ADC.

### Principal Features of the Model

A mathematical model was developed and programmed in FORTRAN IV for the IBM 7044 computer at RAND. The model solves the following kind of optimization problem: it finds the least-cost WRS kits that will keep the probability of a stockout after $K$ cannibalizations less than or equal to some target objective $\alpha$. The user specifies the number of allowable cannibalizations, $K$, and the level of protection that the kit is supposed to provide, $\alpha$. For instance, if we set $K=1$ and $\alpha=10$ percent, the model will compute a WRS kit that will provide enough parts support to keep the number of aircraft grounded due to lack of spares at one or less, with a probability of 90 percent.

---

*Now with Bowdoin College

We call the probability of meeting all spares demand the kit's operational rate. This is used as a measure of kit performance. The probability computation considers a number of factors. The first feature of the model is that it takes into account the possibility of utilizing normal base operating assets. It should be pointed out that just because the base has, say, five units of an item, we do not assume it will necessarily have five units on hand at all times. In fact, we compute the probability that 1, 2, 3, 4 or all of these units will either be in maintenance or in the pipeline from the depot. The kit computation then takes these probabilities into account to assess the availability of the normal operating assets of this item in wartime. In order to satisfy subsequent demands, the second feature the model takes into account is the possibility of cannibalizing an aircraft that is already grounded. The third feature is that the model considers the possibility that some parts can be repaired within an aircraft's turn-around time. The fourth feature is related to demand forecast. Just as in the ADC study, past demand data are treated in a Bayesian fashion for predicting future demand. Finally, percentage base repair is treated as an unknown factor to be predicted for kit optimization. Its past data are also treated in a Bayesian fashion similar to the demand prediction.

### Organization

Section II of this paper contains statements of the assumptions on which the model is based, a discussion of the measure of performance and its computation, and a description of the optimization method. In Sec. III, sample data are used to illustrate the effect of considering ordinary base stock levels when computing the costs of war reserve spares requirements for a WRS kit. And in Sec. IV, several possible extensions of the model are discussed.

## II. THE MODEL

### Assumptions

1. The quantity demanded for an item in a fixed period is assumed to follow a stationary probability distribution. In particular, the form of assumed distribution is Geometric (or stuttering) Poisson. This provides a flexibility for modeling the situation in which item demands have a variance-to-mean ratio greater than unity.

2. The quantity demanded for an item is independent of other items. This asssumption is appropriate as applied to a set of recoverable assemblies as long as there are not many complex LRU (Line Replaceable Unit ) module relationships.

3. The wartime demand does not depend on the number of serviceables that might be available from day-to-day operating stocks. In other words, the distributions of quantities demanded in peacetime and in wartime are independent of each other.

4. We assume that there are no holes in the aircraft at the outset of hostilities, i.e., all the aircraft are operationally ready. This probably is not a true picture of the real world, but a crude approximation of an expediting process that takes place whenever an aircraft is grounded for lack of spares.

5. Items to be considered for inclusion in the kit are equally essential for all missions and all such items can be identified.

### Operational Rate, Cannibalizations, and NORS

"Operational rate" is the yardstick used to measure the effectiveness of a WRS kit. The operational rate is defined as the probability that the kit, together with those serviceables available from normal base stocks, can meet spare parts demand during the emergency. What does an operational rate of.

say, 90 percent mean? One statistical interpretation is that if a situation arises requiring use of the kit, then 90 percent of the Air Force units using the kit will not be hampered from performing their mission because they lack spares.

Because of assumption 2, operational rate may be expressed as the product across all items of the individual item operational rates. (The operational rate for a single item is the probability of meeting all demands for that item.) Thus if we let $g_j(k_j)$ be the operational rate for item $j$ when there are $k_j$ units of the item in the kit and base supply, then

$$\prod_j g_j(k_j) \tag{1}$$

gives the overall operational rate. Note that $g_j(k_j)$ is simply the probability that during the war there will be no more than $k_j$ demands for item $j$ in excess of serviceables available from the combined assets of the kit and base supply. Now suppose that there is one aircraft that can be cannibalized for parts, and let $a_j$ be the number of units of item $j$ that are used on one aircraft. Then, in effect, for item $j$ we have $a_j$ units that may be added to the $k_j$ units in the system, so that the probability of meeting all demands on item $j$ has been increased from $g_j(k_j)$ to $g_j(k_j+a_j)$, and the overall operational rate has been increased from expression (1) to

$$\prod_j g_j(k_j+a_j).$$

In fact, in general, if there are $c$ aircraft available for cannibalization, then the effective stock level for item $j$ becomes $k_j+ca_j$, so the operational rate increases from (1) to

$$\prod_j g_j(k_j+ca_j). \tag{2}$$

As it turns out, it is just as easy to optimize the generalized operational rate given by (2) as it is to optimize the special case given by (1).

In the foregoing, we have interpreted (2) as the probability of meeting all demands for spares during the war, given that there are $c$ aircraft available for cannibalization. This interpretation is also the one given in the ADC study. There is, however, a slightly different interpretation of (2) that may be more meaningful. We may rephrase our interpretation of (2) as the probability of meeting all demands for spares without having to cannibalize more than $c$ aircraft. Thus, if we consolidate parts shortages on as few aircraft as possible, we may interpret (2) as the probability that there will be no more than $c$ NORS (*N*ot *O*perationally *R*eady—*S*upply) aircraft.

To compute (2), it is obvious that we need only be able to compute $g_j(k)$ for any $j$ and any $k$, i.e., the operational rate for an individual item $j$ when $k$ units of the item are in the kit and there are no aircraft available for cannibalization.

**Computation of Operational Rate for a Single Item**

In this subsection we describe the computation of the operational rate for a single item, i.e., the probability that the quantity on hand at the beginning of the emergency (WRS plus serviceables from normal base stockage) is sufficient to cover wartime demands. This probability, of course, takes into account the uncertainty associated with the number of wartime demands. It also takes into account the uncertainty associated with other relevant quantities listed below.

1. The amount of on-hand serviceables from normal base operations (exclusive of WRS).
2. Wartime and peacetime demand rates.

3. Percentage base repair.

For the sake of definiteness in what follows, let us assume that we are talking about an item that can be repaired during the war, at base level, and within the turn-around time of the aircraft. We assume, for now, that the wartime demand has a Poisson distribution. We begin by assuming that expected wartime demand, percentage base repair, and the amount of on-hand serviceables from normal base stocks are, contrary to what was said above, known quantities denoted by $\lambda$, $\rho$, and $X$, respectively. Let $k$ denote the quantity of the item in the WRS kit. Then the probability that all wartime demands for the item can be met via either the kit, repair, or on-hand peacetime serviceables is simply

$$F[k+X;\ \lambda(1-\rho)], \tag{3}$$

where $F[\cdot;\ \nu]$ denotes the cumulative Poisson distribution function with mean $\nu$:

$$F[y;\nu]=\sum_{j=0}^{y} e^{-\nu}\nu^j/j!.$$

Unfortunately, as we pointed out before, we do not know $X$. Under suitable assumptions, in particular that peacetime demand has a Poisson distribution, however, we do know its approximate probability distribution [4]. In fact, if we denote average repair time and average resupply time by $R$ and $S$, respectively, so that average response time is $(1-\rho)S+\rho R$, then the probability that $X$ takes on the value $x$ is given by $f[q-x;\ \mu((1-\rho)S+\rho R)]$ for $x>0$, and $1-F[q-1;\ \mu((1-\rho)S+\rho R)]$ for $x=0$, where $q$ is the normal peacetime stock level, $\mu$ is the item's peacetime demand rate, and $f[\cdot;\ \nu]$ is the Poisson density function with mean $\nu$. Thus, if we form the expected value of (3) with respect to $X$, we obtain

$$F[k;\ \lambda(1-\rho)][1-F(q-1;\ \mu((1-\rho)S+\rho R)] \quad +\sum_{x=1}^{q} F[k+x;\ \lambda(-\rho)]f[q-x;\ \mu((1-\rho)S+\rho R)] \tag{4}$$

for the probability of meeting all wartime demands.

In passing from (3) to (4), we have removed the assumption that $X$, the on-hand serviceables from normal stocks, is known. In much the same way, we remove the assumptions that $\lambda$, $\rho$, and $\mu$ are known with certainty. Let us assume instead that our uncertainty and knowledge about $\rho$ may be reflected by $m$ numbers $\rho_1, \ldots, \rho_m$ and $m$ probabilities $r_1, \ldots, r_m$, where $r_i$ represents the probability that $\rho=\rho_i$. Similarly, we assume that our knowledge about $\lambda$ and $\mu$ may be reflected in $n$ pairs of numbers $(\lambda_1, \mu_1), (\lambda_2, \mu_2), \ldots, (\lambda_n, \mu_n)$ and $n$ probabilities $d_1, \ldots, d_n$, where $d_j$ is the probability that $\lambda$ takes on the value $\lambda_j$ and $\mu$ takes on the value $\mu_j$. (The derivation of these numbers and probabilities will be explained below.) Then, if we sum (4) over the different values that $\lambda$, $\rho$, $\mu$ can take on, in each case multiplying it through by the probability that these values will occur, we obtain

$$g(k)=\sum_i r_i \sum_j d_j \Big\{F[k;\ \lambda_j(1-\rho_i)][1-F(q-1;\ \mu_j((1-\rho_i)S+\rho_i R)] \\ +\sum_{x=1}^{q} F[k+x;\ \lambda_j(1-\rho_i)]f[q-x;\ \mu_j((1-\rho_i)S+\rho_i R)]\Big\} \tag{5}$$

as the operational rate for a single item.

It remains to determine the numbers $\rho_i$, $r_i$, $\lambda_j$, $\mu_j$, $d_j$, the probability distributions for percentage base repair and wartime and peacetime demand rates. These are derived by applying the "objective

Bayes" technique [2]. We first describe how the numbers $\rho_1, \ldots, \rho_m$ and $r_1, \ldots, r_m$ are derived (these give the probability distribution for the percentage base repair).

We assume that $\rho$, the percentage base reparable, has an *a priori* probability distribution that we then approximate by a discrete distribution. This approximation is obtained by dividing the interval between 0 and 1 into $m$ equal intervals. We let $\rho_0 = 0$, and let $\rho_i = 0$, be the midpoint of the $i$th interval for $i > 0$. We wish to associate with $\rho_i$ the probability $r_i^0$ that was originally associated with the $i$th interval. (In the case where $i = 0$, $r_i^0$ will be the probability originally associated with 0.) The numbers $r_i^0$ are estimated by simply letting $r_i^0$ be the proportion of items that, within the data period, have had a percentage base repair that falls in the $i$th interval.[1]

For an individual item, the numbers $r_i$ (*a posteriori* probabilities) are obtained by applying Bayes' rule to the item's repair data and the *a priori* probabilities $r_i^0$. Specifically, suppose that during the data period $N$ units of the item have been turned in to base maintenance, but of these $N$ only $K$ units could be repaired on the base. The probability of this event, given the number $N$ of turn-ins and given that the base repair percentage is actually $\rho_i$, is simply

$$\binom{N}{K} \rho_i^K (1-\rho_i)^{N-K}.$$

Hence, by Bayes' rule, the *a posteriori* probability that $\rho = \rho_i$ is given by

$$r_i = \frac{r_i^0 \binom{N}{K} \rho_i^K (1-\rho_i)^{N-K}}{\sum_j r_j^0 \binom{N}{K} \rho_j^K (1-\rho_j)^{N-K}} = \frac{r_i^0 \rho_i^K (1-\rho_i)^{N-K}}{\sum_j r_j^0 \rho_j^K (1-\rho_j)^{N-K}}. \tag{6}$$

We turn now to the determination of the demand rates $\lambda_j$, $\mu$, and their associated probabilities, $d_j$. We assume that $\lambda$ and $\mu$ stand in the relation

$$\lambda = C\mu,$$

where $C$, a constant across all items, reflects the relative level of activity during war and peacetime. (We have computed $C$ as the ratio of the number of sorties to be flown during the war to the number of sorties per unit time flown during peace.[2]) Thus, if we know the $\mu_j$, we may compute the $\lambda_j$ by the formula

$$\lambda_j = C\mu_j.$$

Finally, the numbers $\mu_j$ and $d_j$ are determined exactly as in the RAND base stockage model [3] (another application of the objective Bayes approach).

The assumption that peacetime and wartime demand have Poisson distributions may be relaxed by assuming some compound Poisson distribution instead. In the work for the Air National Guard, we assumed a geometrically compounded Poisson distribution. This made it necessary to specify an additional piece of data, the variance-to-mean ratio, which we assumed to be the same for all the items.

---

[1] This procedure for determining the number $r_i^0$ is entirely analogous to the use of objective Bayes technique in [2] and [3] except that there it was applied to demand rates rather than percent base repair.

[2] There is an assumption implicit in this statement, viz., demands are correlated to the number of sorties. If this assumption is not appropriate one may have to consider the ratio of total flying hours per unit time or some other appropriate index for measuring the differences in the activity levels of the two situations.

In the foregoing, we assumed that the item whose operational rate we were computing could be repaired, in those cases where repair could be accomplished at all, within the aircraft's turn-around time. For items that cannot be repaired within the turn-around time, the term $\lambda(1-\rho)$ in (3) and subsequent expressions should be replaced by $\lambda$.

**Optimization**

Let $b_j$ be the unit cost of item $j$. Then the cost of the WRS kit will be $\sum_j b_j k_j$, where $k_j$ is the quantity of item $j$ in the kit. The problem of finding the least-cost WRS kit that will have a prespecified operational rate $s$ may then be phrased as follows:

$$\text{Minimize } \sum_j b_j k_j$$

(7)

$$\text{Subject to } \prod_j g_j(k_j + ca_j) \geqq s.$$

It turns out to be more convenient to replace operational rate by its logarithm. When we do this, (7) becomes

$$\text{Minimize } \sum_j b_j k_j$$

(8)

$$\text{Subject to } \sum_j \log g_j(k_j + ca_j) \geqq \log s.$$

This problem may be solved by marginal analysis. We start by setting $k_j = 0$ for each item $j$. We then find an item $j$ for which the ratio

$$\frac{\log g_j(k_j + ca_j + 1) - \log g_j(k_j + ca_j)}{b_j}$$

is a maximum, and increase the corresponding stock level $k_j$ by one unit. We continue in this way until the constraint in (8) (and therefore the constraint in (7)) is just satisfied.[3]

## III. SENSITIVITY ANALYSIS

One of the interesting features of the model described in Sec. II is that it takes account of ordinary base stock levels when computing the war reserve spares requirements. In this section, we use some numerical results to illustrate how such a consideration affects the cost of a WRS kit. For the sensitivity analysis pertaining to other features of the model, such as the ability to incorporate cannibalization into a spares requirement computation and the use of a Bayesian technique for demand prediction, the reader is referred to the classified work we did for ADC and [5].

To examine the impact of base stock levels on the cost of a WRS kit, we have computed some WRS kits by four different methods. The computation is based on data from an F–102 Air National Guard

[3] Strictly speaking, problem (8) is only approximately solved by this procedure. The solution obtained by marginal analysis, however, is "efficient" in the sense that no other policy is better with respect to one criterion (cost or performance) without being worse with respect to the other [5]. It should also be pointed out that when demand rates are large, the functions $g_j$ may not be concave. In this event they should be replaced by their concave majorants as in [3].

Squadron at Boise Airport, Idaho. The data came from 16 aircraft over a 6-month period. During that time the squadron flew a total of 2218 hours and 1299 sorties. In all, we considered 245 items as candidates for the kit. The results of this computation are presented below.

| Method | Cost (In $ thousand) |
|---|---|
| I. Peacetime levels not used during war | 332 |
| II. Peacetime levels used during war, but not in the optimization. | 222 |
| III. Peacetime levels used during war and in optimization | 137 |
| IV. Peacetime levels computed by BSM[a] and used in both war and optimization | 116 |

NOTE: 90-percent performance.
[a] The RAND Base Stockage Model.

Method I assumes that the only supply support during the war would come from the WRS kit. This assumption would be valid were the Air National Guard to operate under a dispersal concept similar to that used by units of the Aerospace Defense Command. In the second computation, it was still assumed that there were no normal operating assets for the purpose of kit optimization. We then came up with a kit that, when evaluated jointly with the normal peacetime assets, resulted in 90-percent performance. This kit cost $222,000. The third method explicitly takes into account the levels at Boise Airport. These base stock levels are a mixture of Chapter 11 stock levels and negotiated levels. A target for performance was again set at 90 percent. The resulting kit cost $137,000.

The meaningful comparison for the Air National Guard is between the second and third kits. By taking peacetime levels into account explicitly when designing a WRS kit, we can effect a cost saving of nearly 40 percent without degrading base supply performance in wartime.

The computation method for the last kit was the same as that for the third kit, except that peacetime levels were computed using the RAND Base Stockage Model. We were interested to see what effect the "efficient" base stock levels in peacetime would have on the cost of a WRS kit. We computed the dollar investment at Boise Airport at its peacetime levels [4] and used the same dollar figure as an investment constraint in maximizing peacetime fill rate. The results indicate that the peacetime levels computed by the Base Stockage Model seem to have provided a somewhat better set of levels with which to start the war reserve computation.

These numerical results indicate that if we could explicitly take account of the availability of peacetime assets to support wartime operations when we design a WRS kit, we could either substantially reduce costs or substantially increase supply performance.

[4] This figure was estimated to be about $538,000.

## IV. EXTENSIONS OF THE MODEL

In this section, we briefly discuss two possible extensions of the model.

### Joint Optimization of Peacetime and WRS Stock Levels

The model we have described here sets only the WRS levels; it treats peacetime stock levels as externally determined data. In other RAND stockage work, models have been developed for determining the peacetime levels. These models (as well as current Air Force policy) ignore the possible role of peacetime stocks in augmenting WRS. It might be possible to derive more cost-effective stockage policies by optimizing both the WRS and the wartime levels at the same time in order to meet constraints on *both* the peacetime and the wartime performance of the combined stockage policy. The theory for such a joint optimization has already been developed [1].

### Use of NORS as a Performance Criterion

In Sec. III we pointed out a connection between operational rates for various levels of cannibalization and the probability distribution for the number of NORS aircraft. Since expected number of NORS is perhaps a somewhat more meaningful performance measure than operational rate, one might well ask why we do not optimize expected NORS rather than operational rate. Unfortunately, the expression for expected NORS is not nearly as tractable mathematically as is operational rate. In fact, the marginal analysis technique of the preceding section is inapplicable, at least directly, to the problem of optimizing expected NORS. Thus, how to optimize expected NORS directly is a topic of current research. Fortunately, however, kits that give high operational rates tend to give low expected NORS, so that when we optimize operational rate, we are, to a large extend, optimizing expected NORS.

## REFERENCES

[1] Brooks, R. B. S., *Some Linear Programming Applications to Stockage Problems*, The RAND Corporation, RM-5422-PR (September 1967).

[2] Feeney, G. J. and C. C. Sherbrooke, *An Objective Bayes Approach for Inventory Systems*, The RAND Corporation, RM-4362-PR (March 1965).

[3] ———, *A System Approach to Base Stockage of Recoverable Items*, The RAND Corporation, RM-4720-PR (December 1965).

[4] ———, *The* $(s-1, s)$ *Inventory Policy Under Compound Poisson Demand: A Theory of Recoverable Item Stockage*, The RAND Corporation, RM-4176-PR (March 1966).

[5] Fox, B., "Discrete Optimization via Marginal Analysis," *Management Science* (November 1966).

* * *

# A COST-BENEFIT ANALYSIS OF MILITARY AIRCRAFT REPLACEMENT POLICIES

A. James Boness and Arnold N. Schwartz[1]

*Institute of Naval Studies, University of Rochester*

**ABSTRACT**

This paper describes a method of solving aircraft service life problems. The particular application concerns aircraft in the Naval Advanced Jet Training Command. The method of solution is comparative present value analysis of alternative replacement policies. The likely risks of estimation errors are reflected in the comparisons of present values. Differences are noted in the benefits associated with each policy, but external to Naval Aviation. Since the values of these benefits can be determined only at a higher level of decision-making, the result of the study is not a conclusive selection among policies, but a schedule of present values on the basis of which, together with values of the external benefits, a decision can be reached.

This paper discusses replacement policies for aircraft used in the Naval Advanced Jet Pilot Training mission. Taking engineering technology and the training syllabus as given, four feasible plans for introducing replacement aircraft into service are evaluated in terms of the present values of differential costs associated with the plans and in terms of the likely errors in cost estimates used in calculation of the present values. The trade-off between present value of costs and planning flexibility is emphasized in choosing a recommended time pattern of aircraft replacement. The specific aircraft mixes considered are the TF–9J/TAF–9J and the TA–4F/A–4B. The first is the currently employed mix; the second is the proposed replacement.

The problem is to select an optimal time-pattern of replacement of F–9's by A–4's, given technological differences favoring the A–4 and increasing costs of maintaining squadrons of F–9's. Replacements by aircraft types other than the A–4 are considered impractical. Four feasible plans for introducing A–4's through a 5-year period are evaluated in terms of current best estimates of the related costs of the plans and in terms of the flexibility of modifying each plan given future better information concerning the relevant costs. The method of analysis is comparative present value of expected costs.[2]

## I. AN ECONOMIC SERVICE LIFE FOR THE F–9J

The F–9J was introduced in 1954. Deliveries from new production were accepted until 1959, totalling 895. In their current designations, as of 30 June 1967, the operating inventory contains 298 TF–9J's (two-seated version), 47 AF–9J's, and 74 TAF–9J's (both one-seated versions).

---

[1] This paper is a revision of INS RC 23, "Interim Report on the Assignment of Aircraft to the Naval Advanced Jet Pilot Training Mission" conducted under contract N00014–14–68–A–0091 at the Institute of Naval Studies, Center for Naval Analyses of the University of Rochester and dated November 1967. The authors are economic analysts at that organization, the former currently on leave-of-absence as visiting associate professor of finance at the State University of New York at Buffalo. No part of this paper necessarily represents the opinion of the Department of the Navy.

[2] A dynamic programming solution for jointly optimal replacement rates and maintenance/overhaul cycles was also employed in this specific situation. Programming routines may be particularly sensitive to errors in estimations ofcosts and insensitive to possibilities for flexible, on-going management of any proposed maintenance and replacement policies. Ad hoc present value methods, as presented in this paper, and alternative mathematical programming methods both tend to neglect the likelihood of managerially important information occuring subsequent to the specific management decision. But the *non*-programmed solution may be more amenable to subsequent interpretation and reevaluation.

The F–9J is used in the advanced jet phase of the pilot training program. Six squadrons of approximately 60 aircraft each are assigned to training. In addition, there are a few undergoing conversion to drone capability, a few assigned to operating units as utility aircraft or to Research, Development, Test, and Evaluation (RDT&E), and a few temporarily in storage.

Initially the planned service life of the F–9J was 4,000 flying hours. Subsequently the service life was extended to 6,000 hours, after engineering changes and modification costs. The distributions by age and by accumulated flying hours, as of 30 June 1967, of one-seat and two-seat F–9J's are given in Table 1.

The immediate consideration is whether to extend the planned service life of the F–9J to 8,000 flying hours. A more general statement of the problem is to decide, for current planning and budgeting purposes, on the dates, rates, and conditions under which to retire F–9J aircraft and replace them with A–4 type aircraft. We take the mission of the F–9J aircraft as given and do not consider its usefulness for any other purpose. We also assume that the only practical alternative to the F–9J is some mix of TA–4F and A–4B aircraft. Finally, we take the advanced jet phase training program as given and do not consider trade-offs against the A–4 Combat Replacement Air Wing (CRAW) program or against the basic phase of pilot training. These assumptions are eminently realistic, and their net effect is to bias the results in favor of retaining the F–9J beyond a 6,000-hour service life.[3] The age and flying hour distributions of TA–4F and A–4B aircraft are given in Table 2.

This narrows the problem to a question of determining how long the F–9J should be retained in active inventory while the A–4 is available and capable as a replacement. Immediate economic considerations include the cost of a given buy of TA–4F's and the overhaul costs of restoring F–9J's to extended service lives of 8,000 hours. Additional considerations include the operating, maintenance, and support costs, both fixed and variable, coincidental to the extended service lives of the F–9J and to each of the alternative mixes of A–4 replacements. We include attrition losses in the operating cost. These costs are to be estimated over a 4-year period beginning with Fiscal Year (FY) 1969, (FY–69 ends on June 30, 1969). We collapse all benefits and costs associated with an advanced jet pilot training program which are expected to be realized after FY–1972 into residual values as of the beginning of FY–1973. All costs and residual values are discounted at a 10 percent annual rate.[4]

The residual value is the larger of two estimates. The first estimate is the value of retiring the aircraft: this may be negative if removal costs are greater than the scrap value; it may be a nominal transfer price in selling the aircraft to the Military Assistance Program (MAP); or it may be the sum of salvage values of components common to other still-operating models of aircraft.[5]

The second estimate is the value of retaining the aircraft in active inventory. It is the difference between the alternative costs of overhauls or modifications sufficient to sustain the aircraft life over a particular additional number of months and the costs of substituting "used" aircraft of ages such that they could perform the same mission with the same reliability over the same additional number of months, but no longer. In the present context, the used or second-hand aircraft is the A–4B, in part, which after its own modification can be substituted for the TAF–9J. The "cost" of the modified A–4B is the sum of the cost of its modification and its value in its best alternative use without modification.

---

[3] Thus, when we find that the A–4 alternative should be chosen, we do so even while underestimating the value of this alternative. This is due to external benefits of employing A–4's, particularly in subsequent CRAW training. Similarly, we ignore technological and morale considerations which favor an early introduction of the A–4's, but are difficult to measure.

[4] This "social cost of capital" had been administratively determined and is subject to questioning. If it is too high a rate, its effect is to bias the choice of replacement rates against near-term replacement.

[5] A discussion of relevant issues in determining residual values is given in Appendix A.

TABLE 1. *Ages and Flying Hours of Current Aircraft as of 1 July 1967*

| TF–9J's | | | |
|---|---|---|---|
| Age (in months) | | Flying hours | |
| Range: 79 to 120 months with 298 entries | | Range: 1276 to 5416 hours with 298 entries | |
| Range | Entries | Range | Entries |
| 79–87 | 123 | 1200–1699 | 11 |
| 88–96 | 70 | 1700–2199 | 19 |
| 97–105 | 48 | 2200–2699 | 71 |
| 106–114 | 48 | 2700–3199 | 69 |
| 115–123 | 9 | 3200–3699 | 26 |
| ..... | ..... | 3700–4199 | 19 |
| ..... | ..... | 4200–4699 | 21 |
| ..... | ..... | 4700–5199 | 48 |
| ..... | ..... | 5200–5699 | 14 |

| TAF–9J's | | | |
|---|---|---|---|
| Age (in months) | | Flying hours | |
| Range: 126 to 157 months with 74 entries | | Range: 329 to 3746 hours with 74 entries | |
| Range | Entries | Range | Entries |
| 126–130 | 3 | 300–799 | 5 |
| 131–135 | 3 | 800–1299 | 5 |
| 136–140 | 4 | 1300–1799 | 10 |
| 141–145 | 15 | 1800–2299 | 26 |
| 146–150 | 16 | 2300–2799 | 19 |
| 151–155 | 29 | 2800–3299 | 5 |
| 156–160 | 4 | 3300–3799 | 4 |

| AF–9J's | | | |
|---|---|---|---|
| Age (in months) | | Flying hours | |
| Range: 140 to 155 months with 47 entries | | Range: 1276 to 4328 hours with 47 entries | |
| Range | Entries | Range | Entries |
| 140–143 | 5 | 1200–1599 | 4 |
| 144–147 | 12 | 1600–1999 | 7 |
| 148–151 | 14 | 2000–2399 | 8 |
| 152–155 | 16 | 2400–2799 | 12 |
| ..... | ..... | 2800–3199 | 7 |
| ..... | ..... | 3200–3599 | 7 |
| ..... | ..... | 3600–3999 | 1 |
| ..... | ..... | 4000–4399 | 1 |

TABLE 2. *Ages and Flying Hours of Proposed Aircraft as of 1 July 1967*

| TA–4F's | | | |
|---|---|---|---|
| Age (in months) | | Flying hours | |
| Range: 0 to 25 months with 137 entries | | Range: 0 to 587 hours with 137 entries | |
| Range | Entries | Range | Entries |
| 0–3 | 36 | 0–74 | 45 |
| 4–7 | 36 | 75–149 | 19 |
| 8–11 | 36 | 150–224 | 14 |
| 12–15 | 25 | 225–299 | 19 |
| 16–19 | 1 | 300–374 | 20 |
| 20–23 | 1 | 375–449 | 10 |
| 24–27 | 2 | 450–524 | 8 |
| | | 525–599 | 2 |

| A–4B's | | | |
|---|---|---|---|
| Age (in months) | | Flying hours | |
| Range: 95 to 130 months with 319 entries | | Range: 928 to 3881 hours with 319 entries | |
| Range | Entries | Range | Entries |
| 95–99 | 70 | 900–1199 | 1 |
| 100–104 | 88 | 1200–1499 | 6 |
| 105–109 | 82 | 1500–1799 | 19 |
| 110–114 | 57 | 1800–2099 | 44 |
| 115–119 | 20 | 2100–2399 | 44 |
| 120–124 | 1 | 2400–2699 | 67 |
| 125–129 | | 2700–2999 | 69 |
| 130–134 | 0 | 3000–3299 | 38 |
| ..... | 1 | 3300–3599 | 20 |
| ..... | ..... | 3600–3899 | 10 |
| ..... | ..... | 3900–4299 | 1 |

The decision rule for choosing a date or the dates after which F–9J's will be phased-out of the operating inventory is simple, but not, perhaps, self-evident. The rule is, other things equal, to select that date corresponding to the least expected cost transition, where expected costs are measured in terms of present values as of 1 July 1969 and include all relevant operating, maintenance, and support expenditures.

In the absence of a statistically significant least-cost transition date, the transition plan corresponding to the greatest command flexibility is preferred. Costs are estimated, where possible, on a basis of statistical analysis of data gathered from past experience in operating F–9J's and A–4 series aircraft. Possible delivery schedules of new TA–4J's are assumed to be 0, 60, or 120 per year.

## II. AIRCRAFT MIXES FOR SUSTAINING THE ADVANCED JET TRAINING PROGRAM

This section describes and compares the four plans chosen for detailed evaluation; it also explains the basis for choice among a larger family of possible plans. The results of our investigation are summarized in Table 3. The detailed evaluations resulting in the numerical entries in this table are described in Sec. III.

TABLE 3. *Summary of Present Values of Costs Associated with Plan A, B, C, and D*
(Millions of dollars)

| Services | Costs | | | |
|---|---|---|---|---|
| | Plan A | Plan B | Plan C | Plan D |
| Acquisition | ..... | $244.0 | $212.6 | $121.7 |
| Contract cancellation | $2.0 | ..... | ..... | 1.0 |
| NARF expansion | 2.5 | ..... | ..... | 0.5 |
| PAR/overhaul/conversions | 105.6 | 10.2 | 24.3 | 83.8 |
| Engine reworks | 9.9 | 16.9 | 12.1 | 11.3 |
| Intermediate and organizational maintenance and support | 114.2 | 128.0 | 116.7 | 115.0 |
| Subtotal | 234.2 | 399.1 | 365.7 | 333.3 |
| Residual values | | | | |
| F–9 | 4.2 | 11.8 | 10.7 | 4.0 |
| A–4 | 10.0 | 111.0 | 117.8 | 84.6 |
| Net sum | $220.0 | $276.3 | $237.2 | $244.7 |

Plan A is an extreme. It delays introduction of A–4 training squadrons until after 1972, by which time attrition will have reduced the F–9 capability enough to jeopardize the programmed flight hours of the training program. Plan B is another extreme: it accelerates the rate of introduction of A–4's by introducing one squadron before FY–1969, three more squadrons in FY–1969 on a buy of 120 new TAF's, and the final two squadrons in FY–1970 on another buy of 120 TA–4F's. The more moderate plan C begins as plan B, but accelerates more slowly, buying 60 new aircraft in each of FY–1969, 1970, 1971, and 1972. Plan D is a compromise between plans A and C: it begins as plan C by introducing one squadron before FY–1969, procures an additional 60 TA–4F's in FY–1969, but then postpones additional procurement until FY–1972, when it procures 60 new aircraft.

We thus have four plans, including the two extremes of postponed and accelerated transition from F–9's to the A–4's; the intermediate plans were chosen sequentially, in search for the least-cost plan after the boundaries of costs of the extreme plans were derived. The search was discontinued because comparisons of the costs of the already evaluated plans showed small variability relative to the probable sizes of estimating errors in preparing the cost estimate. Plan C is recommended on the basis of its cost advantage over plans B and D and, with respect to plan A, its relative flexibility, which would insure satisfaction of the training program requirements in the event of adverse developments, insurance that is lacking in plan A.

**Flexibility and Feasibility of the Plans**

"Flexibility" refers to the short-run ability of a plan to respond to changes in the programmed pilot throughput rate, in utilization rates, in policy on the mix of two- and one-seated aircraft in the squadrons, and in the overhaul program within the Naval Aircraft Rework Facilities. "Feasibility" refers to those plans which, with respect to both the depot maintenance program and the pilot throughput program would probably work. All of these plans are feasible. Plan A is the least flexible. The likely changes that might test the flexibility of the plans are discussed separately below.

*The mix of two-seated and one-seated aircraft.* The likely effects of different combinations of single- and double-seated aircraft of either A–4 or F–9 types on the rates of aircraft utilization and on the elapsed time for completion of the training program need to be evaluated. Two-seated aircraft seem to have been preferred, for they constitute 60 percent of the aircraft assigned to the training squadrons and have had higher utilization rates; however, the mix of two- and one-seated aircraft is variable, subject to the demands of the training schedule. Some of the training flights must be in two-seated aircraft; others, in fact, can be flown in either two- or one-seated aircraft. It is important to assess the possibility of varying the mix at this time, since the total cost of the A–4 options depend heavily on the number of two-seated TA–4F's which would be bought.

The only apparent objection to having a higher percentage of single-seated aircraft is a weather factor. Both one- and two-seated aircraft can be used in good weather, which is about 80 percent of the time at the established training bases. In foul weather, the instructor pilot can take off and land a two-seated aircraft, allowing the trainee pilot to complete his in-flight practice, whereas in single-seated aircraft, the trainee would not yet be competent enough to solo under instrument weather conditions. It can be easily shown, however, that changing from a 60/40 mix to a 50/50 mix reduces total weather-availability from 0.80 to no less than 0.75, even assuming that single-seated aircraft cannot fly 50 percent of the time while two-seated aircraft fly independently of the weather. Furthermore, there is some flexibility in scheduling training flights and, indeed, in substituting ground instruction and in-flight training for each other according to weather conditions.

In addition, the average monthly rates of utilization are composites of one- and two-seated rates. The maximum rate of utilization is determined solely by maintenance, support, and supply conditions. The differentiation of one-seaters from two-seaters, except for the weather-factor argument, seems unfounded. In view of weather factors, one-seaters should be able to fly 80 percent as much as two-seaters. Thus, the composite average utilization rate for two-seaters realized in FY–1967, 47.2 flight hours per month, consisted of weighted proportions of 55 hours for the TF–9J and 36 hours for the TAF–9J. The one-seaters flew only 65 percent as much. Presumably, the one-seater is under-utilized because of command preference for the two-seater. Since such command preference is not indulged by a higher command which determines the training syllabus, we conclude that it (the preference) is not motivated by factors relevant to this analysis.

*Relative "effectiveness" of F–9's and A–4's.* The final preliminary matter concerns possible changes in the training schedule as they pertain to the differences between F–9's and A–4's. The principle problem is the differences in engines of TA–4F's and of A–4B/C's. At most, this difference would lengthen the schedule by 1 week and several flight hours, perhaps 8 hours per student. On the other hand, the students would have learned about operating two engines instead of one, a quality difference which would be most directly appreciated by a saving in transition time for those trainees who continue their specialization in the A–4. Moreover, the rate of climb of the A–4 is such that the weaponry portion of

the training schedule could be reduced by about 4 hours. So the disadvantage of the A–4 amounts to an increase in training time of about 5 percent, most of which is compensated for by the two advantages of the A–4. In fact, previous procurement of TA–4F's for the training command suggests that the advantages out-weigh the disadvantages. All other existing disadvantages would be corrected for in a modernization program attached to the introduction of A–4B's into the training command. The cost of this modernization program is included in the subsequent analysis.

*Attrition losses of aircraft.* The expected attrition rate of F–9 aircraft used in Navy planning is one loss per 10,000 flight hours. The actually experienced rate during FY–1963–67 has been 0.75 loss per 10,000 flight hours.

We apply attition rates of 0.75 per 10,000 flight hours of F–9's and 1.5 per 10,000 flight hours for A–4's to the inventory at the beginning of FY–1969.[6] We do not estimate attrition before that time because of a compensating error in our estimates of available inventory. In particular, we made those estimates on the assumption that F–9's under conversion to drone configurations would be retained and used as drones. This is contrary to current intentions, but the omission almost exactly compensates for the currently expected attrition between the inventory date of our source material and the beginning of FY–1969. The difference, in fact, is that we count just the available F–9 aircraft, whereas the estimated number, counting both drone re-conversion and attrition, is 420; however, the cost of reconversion of drone F–9's must be included in evaluations of the practical alternative service life policies under consideration.

We chose to employ the 0.75 per 10,000 flight hours figure on the following rationale. F–9 losses experienced on 7- to 12-year old aircraft are likely to exceed losses to be experienced on recently overhauled aircraft. On the other hand, there is no clear evidence as yet that attrition experience is a function of age of the aircraft. Although some such effect is likely to exist, it is difficult to measure because of all the other random events which concurrently influence attrition losses. Most likely, the quality of the pilot and the demands of sorties in the given mission have a "swamping" effect. Since student pilots in the advanced jet training program are not as competent as most other jet pilots, the attrition rate there would be expected to be higher than average Navy rates. On the other hand, the kinds of flying, carrier qualification and such, are much more abusive in an aircraft assigned to combat replacement training than to one assigned to jet pilot training.

The attrition rate expected for A–4 aircraft once they enter the training activity is estimated by the approximation of actual attrition experience of A–4B's in combat replacement and reserve training. This is 1.5 aircraft per 10,000 flight hours.[7] It biases the results given in Table 3 in favor of plan A if we overstate the attrition rate for A–4's as twice that for F–9's. Losses of TA–4F's under plan C, for example, would be about 20 less at the rate of 0.75 per 10,000 flight hours. This would alter the net present value of the plan by about $15 million and make it less costly than plan A.

*Aircraft utilization rates.* It is important to establish the utilization rates of the alternative mixes of aircraft, so that the required numbers of aircraft for different flight hour programs can be estimated.

---

[6] Elsewhere, attrition is measured as a percentage of total active inventory per month. But for studies such as this, in which different aircraft are considered for the same mission and in which the same aircraft is considered at varying utilization rates, it is necessary to consider the details of the problem more closely. It is possible, for instance, that an unsafe aircraft is flown as little as necessary, so that its per-month attrition experience remains small, while an extremely safe aircraft is preferred for this reason and is flown so much that its monthly rate of attrition is large. The exact nature of the relations among attrition, mission, and time are difficult to derive and, as yet, unknown.

[7] These attrition rate estimates are given by the Naval Air Safety Center.

The basic rule of utilization used in Sec. III is 50 flight hours per month per aircraft. This is for computational ease. In FY–1967 the rate was 47.2 hours per month in the advanced jet training mission. A lower rate might be indicated if a decreasing relation of utilization to increasing age could be discovered. A higher rate such as 54.0 hours per month might equally well be substituted if there is no aging effect and if the single-seated aircraft were utilized as much as the two-seated. A suggested maximum rate of 60 hours per month could be achieved for a short period of time, beyond which changes in maintenance and support policies must be invoked.

Actual average utilizations and their variabilities are given in the document referenced in the footnote on page 1. Among our basic assumptions, we do not distinguish between utilization rates for F–9 and A–4 aircraft. Since the TA–4F's would be brand new, it would seem that they could be utilized at higher rates than A–4B's or any of the F–9's. If true, this would bias our results in favor of plan A.

**Advanced Jet Pilot Flight Hour Programs**

The programmed flight hour program fluctuates to a peak in FY–1970 and then levels off at a lower rate. The projection is approximately equivalent to the training of 1,042 pilots in FY–1969 and FY–1971, 1,134 pilots in FY–1970, and 992 pilots per year subsequently. The actual throughput rate in FY–1968 is 945 pilots. The peak in the program at FY–1970 is an increase of 20 percent over the current level of operation.

The number of flight hours associated with the planned rates of pilot throughput fluctuates between 217,000 and 248,000 per year, averaging about 230,000 hours; however, in separating out the fixed part of total utilization to better appreciate the effect of changes in the pilot throughput rate, we generate lower flight hour programs associated with the same rates of pilot throughput.

A throughput of 945 graduate pilots implies an induction of at least 995 trainees, in view of an expected student attrition rate of 2.5 percent and an assumption that dropouts occur more frequently at the beginning of the jet flying training than near the end of the program. Each pilot flies 144 hours, and since part of the program requires the instructor pilot to fly separately from the trainee, a total of 181 aircraft hours per trainee are needed. This implies a total of 172,855 aircraft hours per year. In addition, the instructor pilots need to fly approximately 113 hours per year each to satisfy qualification and familiarization requirements due to the rotation of instructors into different parts of the syllabus and into the training squadrons themselves. For 225 instructors per year—an average of about 169 at any given time plus a rotational increment of 33 percent per year—these indirect flying hours amount to 25,425 aircraft hours per year. Ignoring ROTC and Reserve flying time, and treating aircraft test time as the final phase of maintenance processes, the total aircraft hours per year associated with an annual throughput rate of 945 graduate jet pilots is 198,280 hours. At 47.2 hours per aircraft per month, which is equivalent to 566 hours per aircraft per year, there are 350 aircraft, exclusive of the maintenance pipeline, needed to support the advanced jet training program at the given rates of utilization and of pilot throughput. For the higher utilization rates of 54 and 60 hours per month, but the same throughput rate, the number of required aircraft are 306 and 275. When we assume a higher throughput rate, we assume that the number of instructors will increase proportionately. Then the low utilization rate requires 420 aircraft, and the higher utilization rates require 367 and 330 aircraft.

Table 4 shows considerable variability, from 275 to 420 aircraft. The upper limit is not practical, since there are barely that many F–9's available with a zero pipeline for logistic support. Furthermore, these estimates of operating aircraft do not compare closely with estimates being derived elsewhere in

the Navy, even though some of the assumptions made are the same. This divergence is due to our more detailed treatment of the flying hour program in terms of direct student flights, instructor flights, and instructor qualification flights. We also have been conservative in our treatment of test time, for example, so that if biased, our estimates are more likely to be biased in the direction of fewer operational requirements than in the direction of overstating the number of aircraft required. The planning ratio used by the Navy is 219 aircraft hours per trainee. Our treatment works out to an average of 209.8 aircraft hours per training graduate.

TABLE 4. *Operating Aircraft Requirements, Exclusive of Pipeline, for Support of the Advanced Jet Training Program*

| Utilization rate (hr./mo.) | Pilot throughput rate (per year) | | | |
|---|---|---|---|---|
| | 945 | 1,134 | 1,042 | 992 |
| | Operating aircraft requirements | | | |
| 47.2 | 350 | 420 | 386 | 367 |
| 54.0 | 306 | 367 | 337 | 321 |
| 60 | 275 | 330 | 303 | 288 |

**Repair Workloads and Capacities**

We consider here the feasibility of the plans, plan A in particular, in terms of the implications of the mixes with respect to workloads and capabilities at Navy Aircraft repair facilities. Coincidentally, we generate pipeline estimates for the options listed in Table 5.

TABLE 5. *Estimated Quarterly Workload for Depot Rework, F–9J*

| Options | Number of aircraft | | | |
|---|---|---|---|---|
| | PAR | Overhaul | Pipeline | Pipeline alternate [a] |
| High | 65 | 15 | 42 | 68 |
| Low | 50 | 10 | 28 | 46 |
| NASC plan | ..... | 53 | 42–47 | 68–76 |

[a] Estimated on basis of statistical analysis of current data.

The following discussion focuses on the depot maintenance of F–9 service lives beyond 6000 flight hours. It is assumed that the F–9 rework and overhaul programs would be done at the Pensacola Naval Aircraft Repair Facility. It is further assumed that necessary modification of existing A–4B's would be accomplished by a contractor who would deliver new TA–4's, that the Progressive Aircraft Rework (PAR) program on A–4 aircraft would be done at the Jacksonville facility, and that the extent of that program would be easily manageable, at least initially, relative to the PAR and overhaul programs for the F–9's.

First, we must update the inventory of F–9 aircraft to 1 July 1969. This is easily accomplished by assuming an approximate monthly flight hour rate of 50 hours per aircraft and by adding 1200 flight hours (24 x 50) to each group of aircraft described in Table 1. As a result of this adjustment, as of 1 July 1969, all of the single-seated F–9's, 121 of them neglecting attrition, will be more than 14 years old, although none of them will have accumulated more than 4750 total flight hours. The two-seated TF–9's, however, will have accumulated considerably more flight hours. Approximately 60 of these will have exceeded 5000 flight hours and thereby will have become "eligible" for overhaul during the period preceding 1 July 1969. During the next year, an additional 45 will require overhaul. During FY–72 and FY–73, about 100 more overhauls will be required. Subsequently, the workload will diminish. But by these later years the single-seated F–9's will be accumulating over 5000 total flight hours and entering the overhaul program.

Of course, selectivity in scheduling both flying hours and overhauls for specific aircraft allows some flexibility in planning the repair and overhaul workload; however, considerations such as those made above require the rework facility to be prepared to overhaul F–9's at an annual rate fluctuating between 45 and 60 aircraft. During the last half of FY–68, there then would be some slack for completion of the learning process and improvement of overhaul techniques, delivery of necessary support equipment and spare part inventories, and so on. It is questionable whether the lead times for ordering parts and equipment will be sufficient to permit overhauls at an annual rate of 40 aircraft beginning on 1 July 1968.

In addition, the recurrent Progressive Aircraft Rework process must be applied to these F–9's not undergoing overhaul and to those early overhauls which will be due for PAR after the average interval of 18 months. Table 5 gives low and high estimates of the rework workload on the assumption that F–9's will continue to be fully employed in the advanced jet training activity.

In this table, the high workload is associated with a total operating inventory of 400 aircraft, the low workload is associated with an inventory of 300 aircraft and the Naval Air Systems Command plan as described below. These limits are consistent with the alternative plans shown in Table 3. Pipeline for the high workload is derived on the assumption that PAR's average 40 calendar days and overhauls average 80 calendar days. For the low workload, the corresponding assumptions are 35 days and 71 days, respectively. The assumption that 18 months is an appropriate PAR cycle may be questionable when dealing with aircraft that are 10 to 15 years old. Whether the low workload assumptions are realistic is considered elsewhere [Appendix B, INS document cited in first footnote]. But it is already apparent that an investigation must be made to extend the capacity of the Pensacola plant to accommodate a more than doubling of its facilities for F–9 reworks. There may not be floor space enough to continue the other activities and concurrently have an average of more than 20 to 25 F–9's in the hangars. An F–9 overhaul program even at the low rate represents an increase in depot maintenance effort equal to more than 30 percent of the current F–9 PAR workload.

The "Naval Air Systems Command Plan" is an alternative rework policy suggested by the Pensacola facility while considering the difficulties of maintaining a PAR process concurrently with an overhaul process. The suggested solution is to discontinue PAR on F–9 aircraft and to substitute for it a 24 month overhaul cycle. But the exact content of the recurrent overhauls has not been determined. It certainly makes no sense to double the depot maintenance effort if all it can accomplish is a 33 percent extension of the interval between depot reworks. It would be more costly and, as shown in the bottom row of Table 5, there would be no reduction in pipeline as a result. The range in the left end of this row encompasses the difference between "high" 80-calendar day overhauls and "low" 71-day overhauls.

To the best of our understanding, man-hours per overhaul are expected to increase over man-hours per PAR about as much as calendar days per overhaul would exceed calendar days per PAR. Recent rework of older aircraft have slightly exceeded 3,000 man hours, varying somewhat with the series of aircraft. The planning norm at Pensacola is now 3,024 man-hours. Therefore, after the early effects of learning and familiarization, each F–9J overhaul is likely to require at least 6,000 man-hours. Material costs, for an extension to 8,000 flight hours, are estimated to exceed $100,000 per aircraft. These are variable costs. In addition, the investment costs incurred in increasing plant capacity and tooling-up for an overhaul program of this size enter into the decision of whether to overhaul the F–9 aircraft.

During the fourth quarter of FY–1967, approximately 800,000 man-hours were applied to reworks of airframes and engines. A 35-percent increase in F–9 workload implies an 8-percent increase in total workload. Furthermore, although the total man-hour norms for the F–9 are in accord with experience over the past 2 years, the average turn-around time in calendar days has been 57 rather than the planning norm of 35. If the estimated planning norm of 71 days for the F–9 overhaul is similarly biased, we have underestimated pipeline in Table 5 by 63 percent. Thus, it is realistic under this hypothesis that the F–9 pipeline would double. The cost of equipment necessary to increase the capacity at Pensacola to handle such a workload is estimated as $2.5 million. In addition, a larger plant might be required, and an expanded labor force of competence equal to that of the existing manning would have to be recruited and trained.

## III. EVALUATIONS AND COMPARISONS

This section is an elaboration on Table 3 above. We now discuss each plan in turn, using plan A as the basis of subsequent comparisons. Comparisons among the different plans are made in terms of relevant cost. Relevant costs are those which differ between the compared plans. For example, material usage and supplies in the operating squadrons may be about the same for F–9 aircraft as for A–4 aircraft. If so, these items are not relevant to this study. If, however, there is a relation between material or supply usage and aircraft age or time since last rework, to this extent these would be relevant costs. Unfortunately, information to establish these relations is scanty. It seems unlikely with respect to supplies. The per-flight-hour estimates of usage do not reflect any underlying relation to aircraft age, nor even to different models-series of aircraft used in the same squadron.

### Evaluation of Options

*Plan A.* Plan A is taken as the base of all following comparisons. This has the effect of treating imputed costs of other plans, not as costs of those plans, but as benefits of Plan A. The residual value of A–4's in Table 3 is an example of this treatment.

Table 6, which is drawn from Tables 1–5, projects the distribution of F–9 inventory flight-hours to the middle of FY–1968, FY–1970, and FY–1972. In making this projection, we consider expected attrition, and we assume an average annual accumulation of 600 flight hours per year. This is somewhat higher than current experience and close to the "moderate" utilization rate considered above. The assumption implies an annual flight-hour program of 240,000 hours, in fact. The approved training program, however, has scheduled a flight-hour program of 247,000 hours in FY–1970. The application of estimated attrition is done on a percentage basis, implying each aircraft is equally likely to be lost in operation. Attrition is also allocated on the assumption that the incidence of (one-seated) TF–9 aircraft attrition will be 1.5 times as great as the incidence of TAF–9 (two-seated) aircraft. The initial figures are rounded for simplicity and in recognition of the approximate nature of the projections.

One implication of the effect of attrition is that by the end of FY–1972 there will scarcely be enough F–9's available to support the advanced jet training program. Replacement aircraft for the F–9's should therefore be planned for and programmed for no later than the first half of FY–1970. This is true if we maintain the jet training program exclusively with F–9's as long as possible.

TABLE 6. *F–9 Inventory Projections by Accumulated Flight Hours*

| Flight Hours | Estimated 12/67 | | Estimated 12/69 | | Estimated 12/71 | |
|---|---|---|---|---|---|---|
| | 2-seat | 1-seat | 2-seat | 1-seat | 2-seat | 1-seat |
| 600 | 0 | 5 | ..... | ..... | ..... | ..... |
| 1,200 | 0 | 5 | ..... | ..... | ..... | ..... |
| 1,800 | 10 | 10 | ..... | 4 | ..... | ..... |
| 2,400 | 20 | 40 | ..... | 4 | ..... | ..... |
| 3,000 | 70 | 30 | 10 | 9 | ..... | 4 |
| 3,600 | 70 | 10 | 19 | 36 | ..... | 4 |
| 4,200 | 25 | 10 | 65 | 27 | 9 | 8 |
| 4,800 | 20 | 5 | 65 | 9 | 18 | 31 |
| 5,400 | 20 | ..... | 24 | 9 | 61 | 23 |
| 6,000 | 65 | ..... | 19 | 4 | 60 | 7 |
| 6,600 | ..... | ..... | 19 | ..... | 22 | 7 |
| 7,200 | ..... | ..... | 60 | ..... | 18 | 4 |
| 7,800 | ..... | ..... | ..... | ..... | 18 | ..... |
| 8,400 | ..... | ..... | ..... | ..... | 56 | ..... |
| Total... | 300 | 115 | 281 | 102 | 262 | 88 |

The costs of such a plan are listed in Table 3. The estimated initial cost for 420 overhauls is approximately $63 million. A preliminary estimate was approximately $71 million, indicating good agreement even though detailed breakdowns are not available; however, some material cost estimates made at Pensacola understate by 183 percent comparable estimates supplied by the vendor. It is also reasonable to question the 6000-man-hour estimate for the airframe overhaul. This is on the assumption that whatever estimating bias that is found in current PAR manhour "norms" will also exist in the norm initially established for overhauls, since the same personnel are making both estimates.

Man-hours required in rework are a function of accumulated flight hours and tour length. The effects of extending tours to 24 months from the current 18 months and the 10 percent increase in accumulated flight hours, from 5,000 to 5,500 on older aircraft utilized at about the annual average for 1 year, suggest that the estimating bias is on the order of 11.65 percent. This represents almost 700 man-hours per rework, or a total of 294,000 man-hours. At a cost of $10 per man-hour, the adjusted total bill for all initial overhauls increases by $2.94 million to about $66 million. The current rate at Pensacola is $10.46, which includes some indirect charges that may increase less than proportionally as a result of the overhaul program.

These costs are expected to occur evenly over a 2-year period beginning 1 July 1968. If so, the present value as of 1 July 1969 is exactly the total amount, $66 million, since the reduction by discounting over the second year is equal to the increment yielded over the first year.

The second series of overhauls or PAR's, on the basis of an estimated 5000 manhour average would cost $52.5 million, with a present value of $37.6 million. To support the increase in workload,

an additional investment equal to 25 percent of estimated replacement value of special equipment used in F–9 PAR's only, seems appropriate as a first approximation. This would amount to an additional $2.5 million.

Finally, among the non-recurring costs of this option, we must consider imputed recovery value in alternative uses. For reasons given later, we tentatively treat the residual value of aircraft as of the end of FY–72 as their "salvage" value or price to the Military Assistance Program. Given that foreign nations would not accept aircraft with more than 8000 accumulated flight hours, there would be about 160 TF–9's and about 70 TAF–9's available at that time.[9] At the current transfer prices to MAP, these aircraft would be worth $4.25 million. This partially offsets total plan A cost. All 230 aircraft at the same time, of course, would be a record-setting and highly unlikely volume for the second-hand market.

There remain two additional values to be imputed in the evaluation of plan A: These are the costs of reconversion of twenty-odd drones, and the values of TA–4F's and A–4B's in uses alternative to the training command. The drone conversion presumably is underway at Norfolk and will be completed before FY–1969; however, its costs are not treated here as sunk costs, i.e., costs which cannot be avoided if plan A or a similar plan is not adopted. The same reasoning applies to the already purchased and delivered TA–4F's; they are of some value to other units if not designated for jet training.

The cost of the drone reconversion is not known and accordingly is taken to be $2.0 million, something like $100,000 per aircraft. This is a "nominal" estimate in the sense that it is necessary to recognize such costs in consideration of alternative plans, but that a more exact determination would probably differ from the amount given here.

The value of A–4 types not introduced into jet training is also inexact at this time. Its final determination requires broader force-level consideration. Tentatively, we assign a value of one-third their acquisition costs to TA–4F's released from designation as training aircraft; we also treat the A–4B's not employed in jet training as excess, both here and in our evaluations of plans B, C, and D. They are valued at their "book" MAP prices. The imputed value of one squadron of released A–4's consisting of 35 TA–4F's and 25 A–4B's is approximately $10 million on this basis.

The final cost of a fixed or capital nature associated with this option is the stand-by cost charged by the contractor for the TA–4F's. This cost is at the rate of $10,000 per day, but the length of the period over which it applies is not clear. Sheer guess-work suggests that this commitment extends to the end of the current fiscal year. If so, the total amount at $10,000 per day is $1.8 million. The present value of this amount as of 1 July 1969 at 10 percent is $1.98 million. Rounding to $2.0 million and tallying all non-recurring costs, the sum is $106.1 million.

The recurring costs of operating and maintaining the F–9's amount to about $185 per flight hour for the TF–9J and $160 per flight hour for the TAF–9J at the intermediate and organizational level. These amounts are taken from Naval sources, but differ in that they do not include pro-rated costs of depot maintenance activities (airframe and engine reworks). The present values of the operating and maintenance cost streams occurring at the intermediate and organizational levels are $114.2 million. The current cost of an engine rework is $4305 each, and the flight hour program will require about 2900 engine reworks through FY–1972. The total dollar cost then is $12.5 million, and, if the reworks occur evenly over the 4-year period, the present value of this amount is $9.9 million. This brings the net present value of the basic plan A to $220.0 million.

[9] Already plan A appears unrealistic: some replacement aircraft must be introduced before the end of FY–1972. On this assumption, the remaining F–9's will be operated up to that time, whereupon they would be available for sale to MAP.

The last type of recurring cost, attrition, is implicit in the calculations which are made concerning the size of the inventory. In the evaluations of the following plans, TA–4F attrition is costed explicitly as a determinant of the TA–4F buy.

Our cost estimates do not include the costs of military labor nor any other common items among the various options. It is assumed that the Maintenance and Operating factors, for example, will be identical or exactly compensating over the period and for the four plans under consideration. This assumption follows from the previous assumption that the alternative type-models of aircraft are to be considered as perfect substitutes in the advanced jet training mission. If so, the A–4B's will be modified, if necessary, to conform with the configuration of the F–9's, and then there should be no serious difference in unit maintenance effort. Nonetheless, any actual differences in the uses of military manpower in the different plans are relevant to the decision.

We also disregard costs of crash damage repair, for both type-model-series options in the various plans. This does not seem important from a cost standpoint. The major effect of this omission is to understate the NARF workload.

*Evaluation of plan B.* Plan B is the accelerated transition to A–4's at the rate of 120 new TA–4F's in each of FY–1970 and FY–1971, with the transition of one squadron of A–4's prior to the beginning of FY–1969. This constitutes a total inventory of TA–4's of 290 aircraft. Because of higher estimated attrition rates, expected losses of A–4's will follow the schedule given in Table 7.

TABLE 7. *Expected A–4 Attrition, Plan B*

| Fiscal year | 60/40 mix | | 50/50 mix | |
|---|---|---|---|---|
| | TA–4F | A–4B | TA–4F | A–4B |
| 1969 | 2 | 2 | 2 | 2 |
| 1970 | 10 | 6 | 8 | 8 |
| 1971 | 23 | 15 | 19 | 19 |
| 1972 | 20 | 14 | 17 | 17 |

The inventory at the end of FY–1972 is thus estimated to be about 235–245 TA–4F's, depending on the accuracy of the attrition rate used, the flying hour program, and the rate of aircraft utilization. This table implies also the following annual inventories of TA–4F's: FY–1969, 103 aircraft; FY–1970, 268 aircraft; FY–1971, 245 aircraft. These are operationally ready aircraft with which squadron personnel have become sufficiently familiar to begin the training program. It is assumed that this familiarization will require 3000 flight hours and 1 month's time. In deriving these inventory estimates, it is also assumed that aircraft will be delivered evenly over the fiscal year. Table 8 summarizes both F–9 and A–4 inventories under plan B. The associated present values of relevant costs are given in Table 3 above. The principal assumptions underlying this and the following tables are that the least-utilized aircraft of each series are assigned to the training program, that average annual utilization of 600 hours will apply regardless of aircraft assignments, and that expected attritions will occur among each aircraft series independent of individual aircraft accumulated utilizations.

The result of these calculations is that one squadron will have been introduced during FY–1968, three more during FY–1969, and the remaining two during FY–1970. Under current planning, a 60/40 mix would require 228 new aircraft and a 50/50 mix would require 192. The TA–4F inventory at the end

TABLE 8. *Training Program Aircraft Inventories Under Plan B with 60/40 A–4 Mix*

| Flight hours | Estimated 12/67 | | Estimated 12/69 | | Estimated 12/71 | |
|---|---|---|---|---|---|---|
| | 2-seat A–4 (F–9) | 1-seat A–4 (F–9) | 2-seat A–4 (F–9) | 1-seat A–4 (F–9) | 2-seat A–4 (F–9) | 1-seat A–4 (F–9) |
| 600 | ..... | (5) | 57 | ..... | 78 | ..... |
| 1200 | ..... | (5) | ..... | ..... | 78 | ..... |
| 1800 | (10) | (10) | 46 | (4) | 49 | ..... |
| 2400 | (20) | (40) | ..... | 7 (4) | ..... | ..... |
| 3000 | (70) | (30) | (10) | 59 (9) | 40 | ..... |
| 3600 | (70) | (10) | (19) | 2 (7) | ..... | 6 |
| 4200 | (25) | (10) | (65) | ..... | ..... | 54 |
| 4800 | (20) | (5) | (65) | ..... | ..... | 93 |
| 5400 | (20) | ..... | (14) | ..... | ..... | 10 |
| 6000 | (65) | ..... | ..... | ..... | ..... | ..... |
| Sub-Total... | (300) | (115) | 103 (173) | 68 (24) | 245 (0) | 163 (0) |
| Total..... | 300 | 115 | 276 | 92 | 245 | 163 |

of FY–1972 therefore contains replacement aircraft for only 2 or 3 years future attrition. We therefore augment plan B to include an additional buy of undertermined size in FY–1972 after better attrition experience becomes available. The size of the buy is irrelevant for present planning purposes and does not influence the results of our comparisons of plans. The cost of procurement would be exactly offset by the residual values of brand new aircraft at the end of FY–1972.

The procurement cost of 240 new TA–4F's over a 2-year period is $216 million at an estimated $900,000 per unit. In addition, support materials will cost about $25 million and spare part procurement for initial stocks will cost about $10 million. These last are not spares used in the system, but are the fixed costs of having inventories on hand so that the maintenance process can operate. Replacements of these stocks will be accounted for below.

Assuming that the sum of $251 million will be paid at the beginnings of FY–1969 and FY–1970, rather than evenly over the 2-year procurement period, the present value as of 1 July 1969 is $239.59 million. These estimates are derived from recent and currently planned procurements as recorded in the estimates for budget submissions. In addition, the A–4B's will need modifications to make them suitable for the training mission and compatible for use with the TA–4F. This cost is estimated at about $20,000 per aircraft, and there will be between 200 and 290 such modifications required. The expected cost of the A–4B program therefore ranges between $4 and $5.8 million. The present value of the total procurement and initiation costs as of 1 July 1969 is estimated as approximately $244 million.

The residual value of the remaining TA–4F's is estimated on the basis of an expected 6000 flight hour life. Similarly, for the residual values of related support equipment and spare part inventories. The present value of this $162.5 million as of 1 July 1969 is $111 million.

Our information on airframe and engine overhaul costs of the A–4 aircraft is incomplete. An earlier study of PAR man-hours on A–4C aircraft suggests that PAR man-hours on the average total 2750. The current price for A–4B PAR's is $31,909, this is an average of the two designated overhaul points, Alameda and Jacksonville. The corresponding price for a TA–4F PAR is $40,340. With an 18-month average tour length, an operating inventory of 380 aircraft would generate 205 PAR's per year during FY–1972. Smaller inventories containing many aircraft would generate estimates of 16 PAR's in FY–

1969, 32 in FY–1970, and 120 in FY–1971. The present value of costs associated with this schedule as of 1 July 1969 is $10.16 million. The rapid phasing-in of A–4's is such that only recently-PAR-processed F–9's need be flown in 1969 and 1970 so that there are no PAR or overhaul costs of F–9's associated with this plan.

Engine reworks for the A–4B currently cost $6,334 and occur on the average after 497 flying hours. Engine reworks for the TA–4F currently cost $8,809 and occur on the average after 385 flying hours. Under the assumption of a 50/50 mix, the present value of the depot maintenance of engines is $16.9 million.

The intermediate and organizational costs associated with F–9 aircraft during 1969, 1970, and 1971 are the respective rates of 5/6, 1/2, and 1/6 the full time costs incurred under plan A. The present value of these costs is $51 million. The intermediate and organizational material costs of A–4 aircraft are not known, but from the relation of direct man-hours per month at comparable utilization rates, we estimate that the increase in lower echelon expenses for the A–4 mix relative to the F–9 is about 5 percent. The present value of lower echelon costs associated with the A–4 mix is then $77 million. The schedule of retirements of F–9's is determined by the planned induction of A–4 squadrons. By using MAP prices and calculating present values, we determine the total net value of F–9's, as of 1 July 1969, to be $5.5 million; however, the "recovery" value of an F–9 retired early enough to contribute to the support of remaining F–9's is $122,500, including engine and net of rework and dismantling costs. By assuming that the first squadron so retired can be employed in this fashion, we increase the present value of residual F–9's to $11.8 million.

The final net present value of estimated cash flows associated with the plan B is $276.3 million.

*Evaluations of other plans.* The other plans may be considered as intermediate mixtures of plans A and B. They involve different rates of transition to A–4 squadrons which lie between the slowest rate, plan A, and the fastest rate, plan B. Plans C and D have been evaluated at the same level of detail and with the same precision as plans A and B. The results are also given in Table 3. The respective aircraft inventories are given in Tables 9 and 10.

Plan C is evaluated with these adjustments: The purchase of TA–4F's at the rate of only 60 per year increases the estimated price to $920,000 from $900,000 per aircraft as in plan B; the slower phasing-in of A–4 squadrons permits cannibalization of 90 F–9's rather than merely 60 as in plan B; attrition of A–4's is less than under plan B; and the average age of the A–4's at the end of FY–1972 is lower than under plan B: These differences establish a greater residual value of the closing inventory than under plan B. Plan C also entails an overhaul of 20 F–9's no later than FY–1970.

Plan D would cost more than A because of higher intermediate and organizational costs for the A–4's, forfeit of the imputed value of one initial squadron of A–4's, retooling costs for production of TA–4's in FY–1972, and the necessity of completing the overhaul program for the F–9's. We also assume a forfeiture of prepayments equal to one-half the forfeiture under plan A. The advantage of plan D would be a better chance of surviving through FY–1972 without a transition to a newer model aircraft. Plan D should be taken seriously only if there is a better aircraft than the TA–4F which will become available soon after FY–1972. In that case, this plan offers more insurance against not fulfilling the planned pilot throughput rates than the questionable plan A and minimizes the increased costs of introducing A–4's incurred in plan B.

Plan D has the disadvantage of concurrent support of small numbers of two different types of aircraft. This disadvantage includes problems in supply support, maintenance skill transferability, and provision for depot level reworks, including Special Support Equipment, management and spare part

TABLE 9. *Training Program Aircraft Inventories under Plan C with 60/40 Mix*

| Flight hours | Estimated 12/67 | | Estimated 12/69 | | Estimated 12/71 | |
|---|---|---|---|---|---|---|
| | 2-seat A-4 (F-9) | 1-seat A-4 (F-9) | 2-seat A-4 (F-9) | 1-seat A-4 (F-9) | 2-seat A-4 (F-9) | 1-seat A-4 (F-9) |
| 600 | ..... | (5) | 29 ..... | ..... | 30 ..... | ..... |
| 1200 | ..... | (5) | ..... | ..... | 30 ..... | ..... |
| 1800 | (10) | (10) | 46 ..... | (4) | 25 ..... | ..... |
| 2400 | (20) | (40) | ..... | (4) | ..... | (4) |
| 3000 | (70) | (30) | (10) | 7 (9) | 40 ..... | (4) |
| 3600 | (70) | (10) | (19) | 43 (35) | ..... | (4) |
| 4200 | (25) | (10) | (65) | ..... | (9) | 6 (8) |
| 4800 | (20) | (5) | (65) | ..... | (18) | 54 (3) |
| 5400 | (20) | ..... | (24) | ..... | (61) | 23 ..... |
| 6000 | (65) | ..... | (19) | ..... | (60) | ..... |
| Sub-Total... | (300) | (115) | 75 (202) | 50 (52) | 125 (148) | 83 (19) |
| Total..... | 300 | 115 | 277 | 102 | 273 | 102 |

TABLE 10. *Training Program Aircraft Inventories under Plan D*

| Flight hours | Estimated 12/67 | | Estimated 12/69 | | Estimated 12/71 | |
|---|---|---|---|---|---|---|
| | 2-seat A-4 (F-9) | 1-seat A-4 (F-9) | 2-seat A-4 (F-9) | 1-seat A-4 (F-9) | 2-seat A-4 (F-9) | 2-seat A-4 (F-9) |
| 600 | ..... | (5) | ..... | ..... | ..... | ..... |
| 1200 | ..... | (5) | ..... | ..... | ..... | ..... |
| 1800 | (10) | (10) | 46 ..... | (4) | ..... | ..... |
| 2400 | (20) | (40) | ..... | (4) | ..... | ..... |
| 3000 | (70) | (30) | (10) | 7 (9) | 40 ..... | (4) |
| 3600 | (70) | (10) | (19) | 14 (36) | ..... | (4) |
| 4200 | (25) | (10) | (65) | (27) | (9) | 6 (8) |
| 4800 | (20) | (5) | (65) | (1) | (18) | 13 (31) |
| 5400 | (20) | ..... | (24) | ..... | (61) | 2 (23) |
| 6000 | (65) | ..... | (19) | ..... | (60) | (1) |
| 6600 | ..... | ..... | (19) | ..... | (22) | ..... |
| 7200 | ..... | ..... | (14) | ..... | (18) | ..... |
| 7800 | ..... | ..... | ..... | ..... | (18) | ..... |
| 8400 | ..... | ..... | ..... | ..... | (12) | ..... |
| Sub-Total... | (300) | (115) | 46 (235) | 21 (81) | 40 (218) | 21 (71) |
| Total..... | 300 | 115 | 281 | 102 | 258 | 92 |

inventories. The same criticism applies to plans B and C, with the exception that here the disadvantages are temporary until completion of the transition rather than long-run.

## IV. CONCLUSIONS

This paper evaluates the economic factors affecting the choice of a replacement policy for specific aircraft in a specific assignment. The method of analysis is derived from capital theory. The basis for choice is the present value of comparative costs of competing replacement options; however, given only moderate differences in comparative costs, each replacement policy is also considered from the point of view of subsequent adaptability in the event of future danger of not fully rendering the flow of services required of aircraft in their assigned task. That is, not only the feasibility of the plan, but also its flexibility to adapt to possible forecasting errors or to future program level changes is considered in relation to the cost of each plan.

Four feasible options are considered, two of which represent the extreme rates of replacement. Two more moderate replacement rates are considered in the other plans, both in order to better evaluate the relation of comparative cost to replacement rate and also to explore the relation of replacement rate to flexibility in administration of the plan with respect to possible errors in estimates. The comparative costs of the plans are given in Table 3, and Tables 6 and 8–10, show the aircraft inventories, distributed according to accumulated usage, at three equidistant points over the given planning period.

These aircraft inventories cannot be used directly to evaluate the administrative flexibility of the plans. They are, nevertheless, our best known way of indirectly indicating the comparative advantages of the plans. At any given time, a flexible plan is one associated with an aircraft inventory which contains some reserve capacity for more intensive usage in the given mission plus some potential for reassignment of the aircraft to other critical missions. Since newer aircraft and two-seated aircraft can be temporarily utilized at a greater rate than older or one-seated aircraft, a plan with relatively more two-seated aircraft and aircraft with lower accumulated flight hours is a more flexible plan. Since, for given accumulated flight hours and provisions for a second flying officer, an A–4 can be re-deployed to alternative service assignments and an F–9 cannot, a plan associated with a relatively large inventory of A–4's is more flexible than one associated with a relatively small number of A–4's.

In general, flexibility varies directly with cost, and the choice of a plan depends on the amounts of administrative flexibility of the plans relative to the costs of the plans. Since there is no direct trade-off between the two,[10] and since the contingencies against which the insurance of administrative flexibility are unknown, the task of cost analysis ends with a summary of present values of relevant costs and indirect evidence of the adaptability of the aircraft inventories associated with each plan.

The dominant interpretation of the cost analysis focuses on the degrees of likelihood that the elements of the forecasts will be realized. For example, we assumed an attrition rate twice as large for the A–4 squadrons as for the F–9 squadrons. If we had equalized these rates at the lower figure, plan B would have become as inexpensive as plan A. Likewise, a 20 percent error in estimating the procurement costs of TA–4F's support equipment and spare parts would make plan B relatively attractive. Under-estimates of A–4 attrition and costs would make plan A still more attractive.

Other critical assumptions underlying the cost analysis which need not be accurate, but which do affect the choice among plans include the following: military manpower applied to maintenance and

[10] Flexibility is insurance against contingent, future opportunity costs of present planning decisions. But since the probabilities of the contingencies can be estimated only at a higher level of decision-making, we cannot calculate an actuarial value of such insurance.

support functions is identical among the aircraft types evaluated; both aircraft type systems are equally supportable, and there are no economies or diseconomies of scale; both aircraft types are capable of the same utilization rates at the same relative levels of intermediate and organizational maintenance effort. We also have taken the training schedule as given and the pilot throughput rates as mandatory.

It appears that plan A is the least costly. But this plan entails the risk of undershooting the planned pilot throughput rate in 1970 and, in any event, of procrastinating into a critical aircraft shortage in the early 1970's. Plan B is probably more expensive than the others and perhaps offers too much flexibility in the administration of the plan, in that it over-protects against the risks of pipeline and attrition losses. Plan C is a little more expensive than plan A, but it thoroughly hedges against the risks of underestimating F–9 pipeline and attrition losses. Plan D hedges somewhat less adequately than plan C against pipeline and attrition underestimates for the F–9, but plan D does not solve the problem of eventually selecting a successor to the F–9, as does plan C. Only if there were a strong likelihood of developing a replacement aircraft superior to the A–4 and to be introduced soon after FY–1972, could plan D have merit. On the other hand, if the A–4 is eventually to be the successor aircraft, the contractor re-start-up costs associated with plan D make it inferior to plan C.

## Appendix A

## RESIDUAL VALUES OF TRAINING AIRCRAFT

The evaluation of the residual value of an asset as of the end of some current planning horizon depends on whether or not a secondhand market will exist for that asset at the time. If no resale market exists, the only alternatives to an asset-holder are to continue to utilize the asset, perhaps in one of several available employments, or to place the asset in storage. The so-called mothball fleet is a monument to this last alternative. The typical reassignment of aircraft type-model-series to successively different missions during the course of their total service lives is an illustration of the first alternative.

If, however, a resale market exists, it is usually more efficient to trade in this market than to incur storage costs. In fact, Kenneth Arrow [1] has recently pointed out that the planning horizon, given a resale market with certain desirable characteristics, need be only one decision period distant. Consistently, one of the salient innovations of defense management in the 1960's has been the development of techniques for evaluating capital budgeting proposals in terms of transfer prices and shadow prices. The economist, when dealing with assets for which no market exists, tends to invent one.

In this report, the resale market is represented by the Military Assistance Program (MAP), through which secondhand aircraft may be "sold" abroad at negotiated prices. Although there are considerable externalities in the MAP program with non-defense and non-economic public interests, MAP maintains a set of transfer prices for existing military hardware. Taking these nominal transfer prices as given (from the point of view of the Navy Department), the alternatives of transferring either F–9 or A–4 aircraft from the training mission to the MAP program were evaluated. The option of transferring either type aircraft to MAP was assumed feasible at the nominal prices.[11] We also assumed that these transfer prices would hold over the 5-year period beginning in July 1967, which served as our planning period. Under all plans considered in the text, we assume transfer of remaining aircraft after

[1] Arrow, Kenneth J., "Optimal Capital Policy with Irreversible Investment," (J. N. Wolfe, Editor) *Value, Capital, and Growth* (Aldine, Chicago, 1968).

[11] We maintain misgivings, however, about the possibility of transferring large numbers of aircraft at these prices. The question, beyond our competence at the time of this study, is whether the MAP price is marginal or average: does it apply to small numbers of aircraft or to numbers of 60-plane training squadrons.

the end of FY–1972 to MAP for series A–4B and all F–9's. The numbers of aircraft sold through MAP varies among plans as noted in section III.

Under plans B and C, some of the F–9's can be cannibalized into inventories of spare parts more effectively than sold to MAP. As new TA–4F's and modified A–4B's are introduced into the training command, older F–9's become excess and available for alternative uses. As long as the replacement rate is not too great, enough of the remaining F–9's will continue in operation to benefit from the spare part support provided by cannibalization. From historical records of spare part usage and from engineering estimates of spare part availability due to cannibalization, the value of spare part inventories generated by cannibalization of retired aircraft has been related to new procurement costs of comparable inventories, net of spare part rehabilitation costs. Since this value exceeds the value of intact F–9's in terms of MAP prices, net of MAP rehabilitation costs, it is used as fully as possible in each of the plans.

The possibility of transfer of F–9 aircraft to missions other than the advanced jet pilot training program does not exist, according to administrative guidance. The only practical reemployment seems to be in the Blue Angels' exhibition and demonstration squadron, and that would require only a small number of aircraft. The benefits of such reemployment would be economically negligible and dominated by either cannibalization values or MAP prices.

Similarly, the possibility of storage of F–9 and A–4B aircraft is considered inferior to any of the preceding uses of such aircraft under any of the plans. Although aircraft do occasionally enter the storage status, this appears to be accidental and short-lived. The option of storage does not seem to enter into moderate- or long-range planning decisions. This difference between aircraft and ship inventory management may be due to the greater rate of technological innovation for aircraft, the relative lack of alternative employments for warships, and/or the nature of military strategic planning.

The remaining type of residual value estimate pertains to the remaining lives of new TA–4F's introduced under plans B and C. In these cases, the values of the remaining service-lives are estimated on a straight-line basis according to the fractions of service lives remaining. This assumes no changes in the relative procurement prices of the aircraft, no changes in the levels of training activity subsequent to FY–1972, and maintenance policies realistically consistent with straight-line use-depreciation of aircraft procurement costs. This treatment is consistent with the availability of several assignments of the TA–4F alternative to the training command. By inspection of Table 3, however, the relative importances of these particular residual values can be appreciated, and errors in the assumed service life of a TA–4F of 6,000 flight hours could easily influence our conclusions stated in section IV.

## Appendix B

## ATTRITION

Aircraft attrition is, of course, of Navy-wide interest. Methods of calculating and forecasting possible attritions vary among individual commands according to local conditions. In general, estimates are made in terms of aircraft losses as a percentage of total aircraft activity or as an expected rate of loss per year per aircraft in active inventory. The rates are sometimes related to measures of aircraft activity, such as time-since-induction into current service tour or accumulated service-life flight hours. The Naval Air Safety Command employs a method based on total accumulated flight hours, and expected attrition rates for planning purposes are estimated principally from historical records of aircraft

losses according to the locations and the kinds of activity the aircraft engage in, by type-model-series of aircraft. For example, aircraft based on a carrier in the Atlantic Fleet have different attrition incidence according to the specific aircraft designs, Atlantic Fleet aircraft have different attrition incidence than Pacific Fleet aircraft, and carrier-based aircraft, because of differences in the conditions of landing and take-off, have different experiences than land-based aircraft. Finally, attrition of aircraft engaged in a pilot training program, other things equal, have different attrition incidence than the same aircraft flown by more experienced pilots.

In general the official attrition rates, calculated in cognizance of these differences, are reliable estimates of future attrition, given aircraft type, mission, and anticipated flight activity. But since attrition losses of aircraft are relatively rare events in the particular non-combat mission with which we are concerned, the reliabilities of the estimates are not uniform.

This is particularly important in considering the assignment of an aircraft such as the A–4B or TA–4F to a new mission for which no historical attrition data exists. One approach might be to consider the attrition incidence of technologically similar aircraft in the given mission, if such comparison is possible. Instead, in this study, we examined the experience of the same aircraft, the A–4B, when employed in a similar mission.

The similar mission currently employing the A–4B is referred to in section I as the Combat Readiness Air Wing (CRAW). This is the part of pilot training next following a specialized syllabus, such as the advanced jet program. In the A–4 CRAW, the pilot develops an expertise in a particular jet aircraft, rather than a general competence in jet aircraft of no direct combat relevance.

Using multiple-variable log-linear regression, we discovered significant, though weak, positive relations of attrition to aircraft age and to the time interval since a particular aircraft had undergone a depot-level maintenance action. We found a stronger relation to such individual aircraft characteristics as total accumulated number of carrier landings and total accumulated flight hours. Since the aircraft under study were to be employed at East-Coast bases, we examined the attrition experience of A–4's in the Atlantic Fleet CRAW.

Without any data base available for an attrition study of the newer TA–4F, we heroically assumed its attrition rate in advanced jet training would be slightly less than that of the A–4B, since some additional safety is attributable to a two-seater over a one-seater. Some confidence was gained by a comparative study of attrition experience of another series of the A–4 type, the A–4C, also employed in CRAW missions. Finally, we rounded upward by one standard deviation the calculated average attrition of A–4B's in CRAW, given the independent variables in our estimating equation, to arrive at the attrition rate used for similar aircraft in advanced jet pilot training. This was indicated by the earlier stage of flying skills characteristic of jet trainees relative to jet training graduates flying in CRAW.

* * *

# A SIMULATED PORT FACILITY IN A THEATRE OF OPERATIONS

Reed E. Davis, Jr.

*Lt. Col., U.S. Army*

and

Robert W. Faulkender

*Major, U.S. Army*

and

William W. Hines

*Georgia Institute of Technology*

**ABSTRACT**

A hypothetical port facility in a theatre of operations is modeled and coded in a special purpose simulation language, for the purpose of conducting simulation experiments on a digital computer. The experiments are conducted to investigate the resource requirements necessary for the reception, discharge, and clearance of supplies at the port. Queue lengths, waiting times, facility utilizations, temporary storage levels, and ship turn-around times are analyzed as functions of transportation and cargo handling resources, using response surface methodology. The resulting response surfaces are revealing in regard to the sensitivity of port operations to transportation resource levels and the characteristics of the port facility's load factor. Two specific conclusions of significant value are derived. First, the simulation experiments clearly show that the standard procedures for determining discharge and clearance capacities take insufficient account of the effects of variability. Second, the response surfaces for ship turn-around times and temporary storage levels indicate that an extremely steep gradient exists as a function of troop levels.

## INTRODUCTION

The primary objective of this paper is to present the results of some simulation experiments designed to investigate the troop requirements necessary for the reception, discharge, and clearance of supplies at a port facility in a theatre of operations. A secondary objective is to illustrate the power of simulation in the solution of transportation and logistics problems of the type considered herein. The magnitude of logistics problems and the inadequacy of standard practice logistics manuals and tables have been obvious in both Korea and Vietnam. The complexity of the logistics planning operation in Korea has been well documented by General Garvin [2] who has vividly described the operation of Pusan Port during the Korean War.

General F. S. Besson, Jr., Commanding General of the U.S. Army Materiel Command, has recently described the nature of the logistics problem in Vietnam for the layman in a popular news magazine [1]. Periods during which inadequate supplies were available for consumption in the field as well as high theft rates have been described in the press.

## A MODEL OF THE PORT

Effective port operations derive from the right combination of an adequate port facility, essential cargo handling equipment, variously trained personnel, and sufficient transportation. The commit-

ment of these resources must provide a balanced capability to receive, discharge, and clear the port of arriving cargo.

The conceptual port considered in these experiments is pictured in Figure 1.

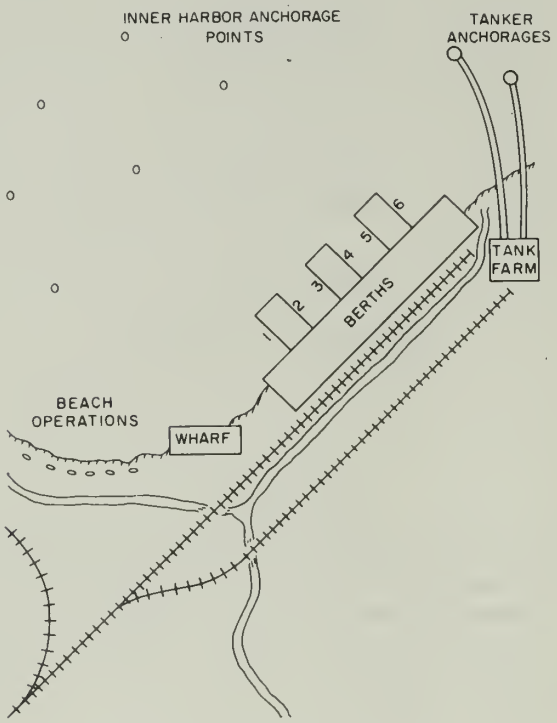



FIGURE 1. Empirical concept of port facility.

The major assumptions employed in this model are noted below:

1. The port facility was established as fixed during the timespan with which the investigation was concerned.[1] See Figure 1.

2. The dry cargo and POL (Petroleum, Oil, and Lubricant) ship interarrival times were assumed to be independent and exponentially distributed on an individual ship basis[2] with constant mean inter-arrival times.

3. No constraints were placed upon troop levels in this study; however, the existence of troop constraints can be easily considered with a slight modification to this model.

4. A constraint on available rail transportation was established at eight trains.

5. A constraint on available transportation for bulk POL was established at three medium truck companies and four trains.

---

[1] This is a reasonable bound for port operations in a *newly activated* theatre of operations. Certainly, port facilities will be improved at some time. An investigation of these improvements is a fruitful area for further study.

[2] Convoy arrivals of varying size could be easily simulated if such an arrival model was appropriate to the problem being studied. In this model, time between dry cargo and POL ship arrivals have the following density functions:

$$\begin{aligned} &\text{Dry Cargo: } f(t) = \lambda_1 e^{-\lambda_1 t}; && t \geq 0 \\ &\qquad\qquad\ =0; && \text{otherwise} \\ &\text{POL: } \quad g(t) = \lambda_2 e^{-\lambda_2 t}; && t \geq 0 \\ &\qquad\qquad\ =0; && \text{otherwise} \end{aligned}$$

6. A constraint on the temporary storage of bulk POL was established at 9800 short tons (the port's POL tank farm capacity).

In this research the results are only valid for the assumed arrival and service rates and model structure as described.

As noted, the inter-arrival times for ships were assumed to be mutually independent, exponentially distributed random variables. Dry cargo ships arrive at a mean inter-arrival time of 30 hours ($\lambda_1 = 1/30$) with a load of 5,600 short tons. POL tankers arrive at a mean inter-arrival time of 80 hours ($\lambda_2 = 1/80$) with a load of 8,400 short tons of bulk POL.

The service times (times to unload ship) were also taken to be mutually independent, exponentially distributed random variables. Mean times for various services were based upon unit capabilities as described below[4]:

1. Light Truck Company—350 short tons per day of cargo (4 tons per truck), based on 75 percent availability of vehicles and two line hauls daily.

2. Medium Truck Company—1,050 short tons per day of cargo (or POL if equipped with petroleum semi-trailers), based on 75 percent availability of vehicles and two line hauls daily.

3. Terminal Service Company—operating on a 20-hour day, two-shift basis; discharges one standard five-hatch ship and loads onto available transportation 700 short tons of cargo daily, or loads onto available transportation 840 short tons of cargo from temporary storage daily.

4. Amphibious Truck Company—transports 700 short tons of cargo daily, based on an availability of 30 vehicles carrying 3 tons per trip and making eight trips per day.

5. Railway Operation Battalion—operates and maintains the eight trains, each of which can transport 700 short tons of cargo (sufficient petroleum tankers are available to equip four trains) daily.

## THE SIMULATION MODEL

The computer programs were written using GPSS III, a special-purpose simulation language provided by the IBM corporation [5, 6, 7]. The language is well suited to the study of problems which can be reasonably viewed as large-scale discrete unit flow and queuing problems. It is a special purpose computer language with its own compiler, allowing the analyst to describe the simulation model in "real world language," thereby shifting a great deal of the translation task to the computer. A GPSS III simulation model is written in terms of the program's 11 entities and their respective attributes [5]. The programmer must simply understand the functions of a set of flow-chart symbols and the rules for combining them. Once the analysis and flow diagram are completed, the program is easily written.

The general flow diagram for the model's dry cargo portion is reflected in Figure 2.

When a cargo ship arrives, it will queue before entering the inner harbor. The inner harbor is tested to determine if its content is less than 12 ships. If space is available the ship will enter the harbor and depart the queue for the inner-harbor. The ship will then test to determine if berth content is less than six (the number of berths available). Based upon the outcome of this test the entering ship will occupy either a berth or an anchorage. In each line the ships will queue and seize (occupy) the service unit(s)[3] necessary for unloading. Upon obtaining the needed service unit(s), seven duplicate "transactions" are created. Each of the duplicate transactions represents 700 tons of dry cargo, whereas the original "transaction" continues to represent the ship and the last 700 ton unit of cargo to be un-

[3] Ships with a berth need only a terminal service company. Ships with an anchorage require a terminal service company and an amphibious truck company.

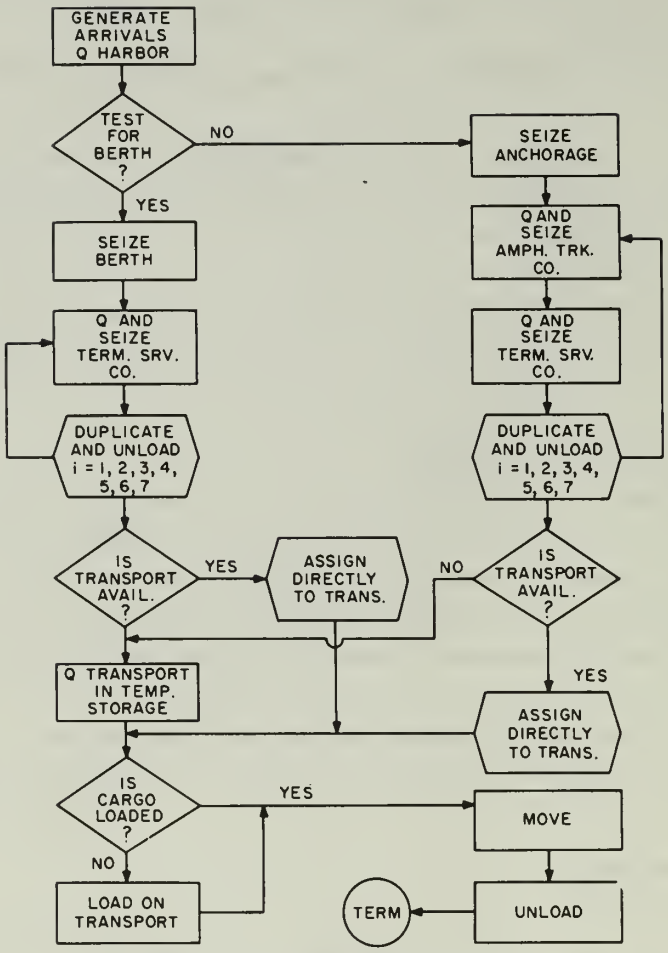



FIGURE 2. General flow diagram (dry cargo).

loaded. The model easily identifies the nature of each transaction by the assignment of appropriate parameter values [5, 6]. The transaction which represents both the ship and cargo is unloaded last.

If transportation is available, the cargo is unloaded from the ship to the transport means, and it is moved to the depot area and unloaded. If transportation is not available, the cargo is unloaded into temporary storage, where it must queue for transportation and a *terminal service company*. When both service units are available the cargo must be loaded, moved to the depot area, and unloaded. As can be seen, the availability of transportation at the time of a ship's unloading is a highly desirable state of successful port operation.

The POL portion is somewhat different in structure from the dry cargo portion. The general flow diagram for POL is reflected in Figure 3. When a tanker arrives it will queue before a POL anchorage and test to determine if its contents are less than two (the inner harbor's POL tanker capacity). On an affirmative to this test the tanker will depart the anchorage queue and seize (occupy) an anchorage. Upon seizure of the anchorage facility, 11 duplicate transactions are created. Of these duplicate transactions, the original one continues to represent the tanker and the last 700 tons of bulk POL to be discharged. Due to the lesser complexity of the POL portion, as compared to the dry cargo portion, separate event chains are created for each type transaction. The duplicate transactions are simply given a preemptive priority which insures that all duplicate transactions are discharged from a tanker prior to

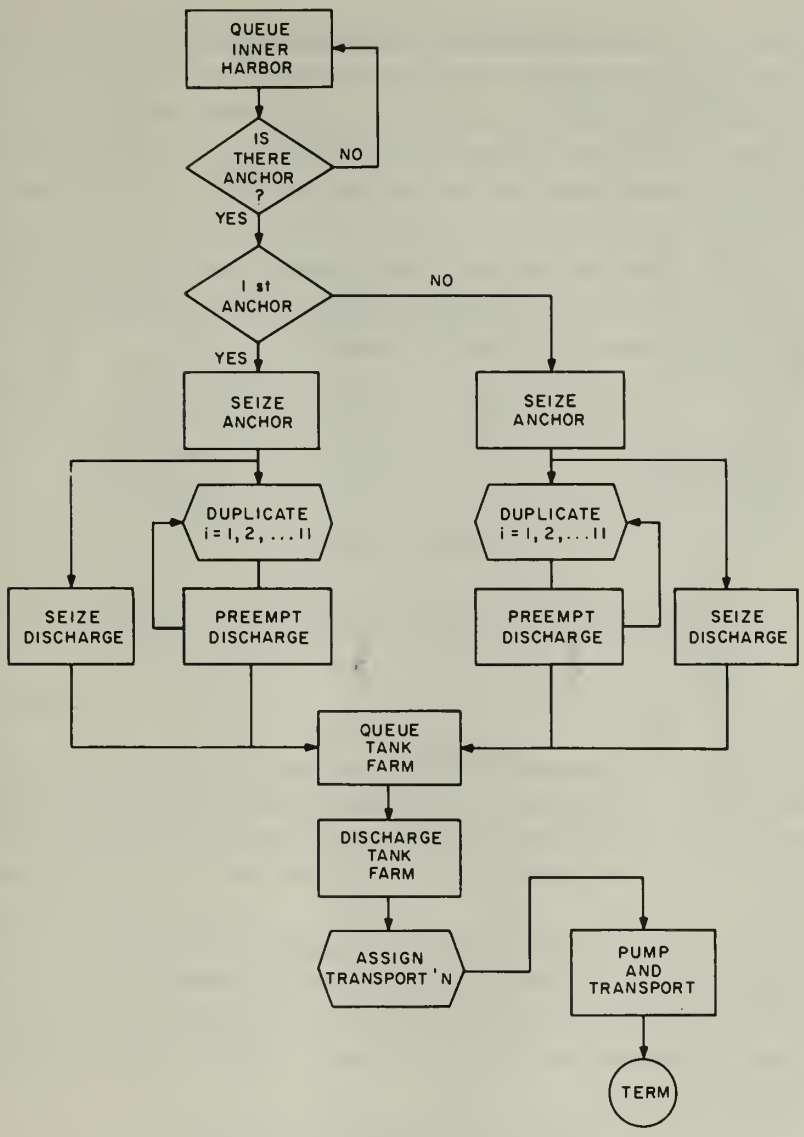



FIGURE 3. General flow diagram (POL).

that transaction which represents the last 700 tons of POL and the tanker itself. All units of POL at an anchorage will queue at the tank farm and test to determine if its contents are less than 14 (tank farm capacity in 700 ton units of bulk POL). With an affirmative result to this test, each 700 ton unit of POL seizes a discharge facility and is pumped into the tank farm with duplicate transactions preceding first.When the last 700-ton unit of POL is discharged, it will cause the release of the anchorage. Each POL unit in the tank farm queues to seize a transportation unit capable of hauling bulk POL. Upon seizure of the transport means, the POL is loaded, transported to the depot area, and unloaded.

## EXPERIMENTAL DESIGN

The number of terminal service companies and transportation companies required to support the daily input were computed, as prescribed in FM 101–10–1, *Staff Officers' Field Manual* [3] to be seven terminal service companies, four light truck companies, two medium truck companies, and two

trains to handle the expected daily input of dry cargo. Additionally, it was determined that three medium truck companies and one train would be required to handle the expected daily input of bulk POL. An initial simulation run was made on this basis with a small margin of allowance for variability of arrivals and service rates. Eight terminal service companies and 13 transportation units (4 light truck, 5 medium truck, and 4 trains) were utilized with this first run. The results reflected that the system could not realize a steady state as the inner-harbor queues for cargo ships and POL tankers grew without bound. This result was anticipated, and simulation was conducted at these levels to determine the time required to saturate the system and demonstrate the fallibility of the prescribed procedures for determining discharge and clearance capacities.

Based on the results of the initial simulation, a two-factor experiment was designed as follows:

| Number of Transportation Units | Number of Terminal Service Companies | | | | | |
|---|---|---|---|---|---|---|
| | 14 | 13 | 12 | 11 | 10 | 9 |
| 31 | A–2 | B–2 | C–2 | D–2 | E–2 | F–2 |
| 33 | A–3 | B–3 | C–3 | D–3 | E–3 | |

## RESULTS

The significant results are reflected in Tables 1 through 5 which show for each experiment (treatment), mean queue lengths, mean waiting times, mean waiting times given a wait $E(w|w > 0)$, major facility utilizations, and aggregate ship and tanker turn-around times, respectively. The most striking aspect of these results is the high sensitivity of the system to the number of transportation units provided.

Based on the information reflected in Tables 1 through 5, experiments A–2 and A–3 are given no further consideration due to the small realization from the additional terminal service company as compared to experiments B–2 and B–3. Experiments D–2, E–2, E–3, and F–2 are given no further consideration due to each presenting queue(s) of importance which are growing without bound indicating instability of the system.

TABLE 1. *Queue Lengths of Importance*

| Experiments | Mean contents | | | | |
|---|---|---|---|---|---|
| | Outer harbor | POL anchorage | Temperature storage | Loading temperature storage | Total storage |
| A–2 | 0.00 | 0.29 | 0.00 | 0.00 | 0.00 |
| A–3 | 0.00 | 0.15 | 0.08 | 0.03 | 0.11 |
| B–2 | 0.00 | 0.27 | 0.04 | 0.04 | 0.08 |
| B–3 | 0.02 | 0.07 | 0.01 | 0.00 | 0.01 |
| C–2 | 0.01 | 0.07 | 0.06 | 0.63 | 0.69 |
| C–3 | 0.01 | 0.18 | 0.06 | 0.52 | 0.58 |
| D–2 | 3.45 | 11.14 | 29.44 | 15.00 | 44.44 |
| D–3 | 0.00 | 0.12 | 0.00 | 0.00 | 0.00 |
| E–2 | 0.00 | 29.70 | 54.31 | 15.64 | 69.95 |
| E–3 | 0.00 | 3.22 | 5.81 | 6.97 | 12.78 |
| F–2 | 18.86 | 18.51 | 37.10 | 17.10 | 54.20 |

TABLE 2. *Waiting Times of Importance*

| Experiments | Meantime/transaction | | | | |
|---|---|---|---|---|---|
| | Outer harbor | POL anchorage | Temperature storage | Loading temperature storage | Total storage |
| A–2 | 0.00 | 16.66 | 0.00 | 0.00 | 0.00 |
| A–3 | 0.00 | 16.47 | 2.39 | 0.79 | 3.28 |
| B–2 | 0.00 | 21.04 | 0.89 | 0.86 | 1.75 |
| B–3 | 0.48 | 6.33 | 0.66 | 0.00 | 0.00 |
| C–2 | 0.24 | 6.41 | 1.40 | 13.79 | 15.19 |
| C–3 | 0.21 | 10.64 | 1.58 | 12.77 | 14.35 |
| D–2 | 94.52 | 873.72 | 115.63 | 60.00 | 175.63 |
| D–3 | 0.00 | 9.29 | 0.00 | 0.00 | 0.00 |
| E–2 | 0.00 | 1969.71 | 236.09 | 75.50 | 311.59 |
| E–3 | 0.00 | 189.82 | 41.77 | 50.13 | 91.90 |
| F–2 | 465.54 | 1588.14 | 184.07 | 90.04 | 274.11 |

TABLE 3. *Waiting Times* ($w|w > 0$) *of Importance*

| Experiments | Meantime/transaction Q | | | | |
|---|---|---|---|---|---|
| | Outer harbor | POL anchorage | Temperature storage | Loading temperature storage | Total storage |
| A–2 | 0.00 | 37.63 | 0.00 | 0.00 | 0.00 |
| A–3 | 0.00 | 31.40 | 8.79 | 5.00 | 13.79 |
| B–2 | 0.00 | 50.50 | 3.27 | 10.00 | 13.27 |
| B–3 | 8.50 | 28.50 | 3.80 | 0.00 | 3.80 |
| C–2 | 11.00 | 28.20 | 4.70 | 24.61 | 29.31 |
| C–3 | 6.00 | 27.09 | 4.54 | 25.17 | 29.71 |
| D–2 | 109.39 | 953.15 | 128.70 | 62.94 | 151.64 |
| D–3 | 0.00 | 28.89 | 0.00 | 0.00 | 0.00 |
| E–2 | 0.00 | 1969.71 | 243.64 | 77.51 | 321.15 |
| E–3 | 0.00 | 285.16 | 54.98 | 54.63 | 109.61 |
| F–2 | 498.09 | 1704.34 | 191.36 | 93.97 | 285.33 |

TABLE 4. *Major Facility Utilization (Storages)*

| Experiments | Fraction of time utilized | | | | | | Time (hours) | |
|---|---|---|---|---|---|---|---|---|
| | Inner harbor | Berth space | POL storage | Term. service company | Transportation | POL anchor | Harbor turn around | POL harbor turn around |
| A–2 | 0.5236 | 0.7996 | 0.7470 | 0.4189 | 0.7664 | 0.7494 | 168.54 | 81.71 |
| A–3 | 0.3933 | 0.6775 | 0.3706 | 0.3581 | 0.5925 | 0.4625 | 169.77 | 66.57 |
| B–2 | 0.4909 | 0.7588 | 0.4737 | 0.4514 | 0.7177 | 0.5011 | 169.15 | 73.04 |
| B–3 | 0.7257 | 0.8395 | 0.2707 | 0.5867 | 0.8278 | 0.3254 | 193.62 | 55.70 |
| C–2 | 0.5738 | 0.7938 | 0.3152 | 0.5069 | 0.6852 | 0.3816 | 194.86 | 63.00 |
| C–3 | 0.5962 | 0.7508 | 0.5522 | 0.5095 | 0.7540 | 0.6457 | 189.14 | 72.83 |
| D–2 | 0.9068 | 0.9938 | 0.9121 | 0.7333 | 0.9989 | 0.8244 | 301.28 | 465.60 |
| D–3 | 0.5141 | 0.8214 | 0.3135 | 0.4127 | 0.6087 | 0.4138 | 179.43 | 61.89 |
| E–2 | 1.0000 | 1.0000 | 0.9994 | 0.6667 | 0.9394 | 1.0000 | 390.75 | 1215.60 |
| E–3 | 0.6650 | 0.8436 | 0.7841 | 0.6046 | 0.9021 | 0.7758 | 216.43 | 115.33 |
| F–2 | 0.9946 | 0.9968 | 0.9897 | 0.5996 | 0.9382 | 0.9680 | 417.31 | 1044.30 |

TABLE 5. *Aggregate Ship Turn-Around*

| Experiments | Time (hours) | |
|---|---|---|
| | Dry goods | POL products |
| A–2 | 168.54 | 98.37 |
| A–3 | 169.77 | 77.04 |
| B–2 | 169.15 | 94.08 |
| B–3 | 194.10 | 62.03 |
| C–2 | 195.10 | 69.41 |
| C–3 | 189.35 | 83.47 |
| D–3 | 395.80 | 1339.32 |
| D–3 | 179.43 | 71.18 |
| E–2 | 390.75 | 3285.31 |
| E–3 | 216.43 | 305.15 |
| F–2 | 882.85 | 2632.44 |

## ANALYSIS OF RESULTS

Of those experiments retained there is little basis for making a choice among them as to the optimal operating policy. Queue lengths and ship turn-around times are acceptable in each case. Of the results portrayed, the points on the three response surfaces (Figures 4–6) are the most revealing. In each case the response surface appears to be relatively uniform when the system is not saturated; however, as saturation of the port's facilities is approached, a significant warp appears in each apparent response surface. The gradient, which each surface displays, indicates that the system is considerably more sensitive to the transportation unit level than it is to the level of terminal service companies. In a qualitative sense, it would appear prudent to select the higher transportation level and to then

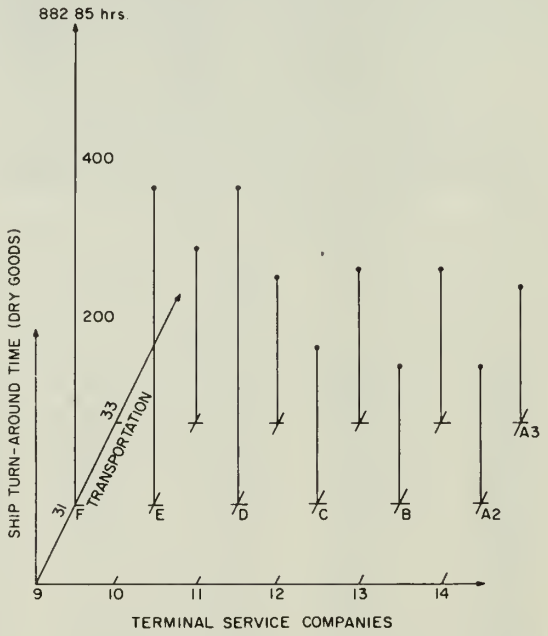



FIGURE 4. Response surface points, cargo ship turn-around.

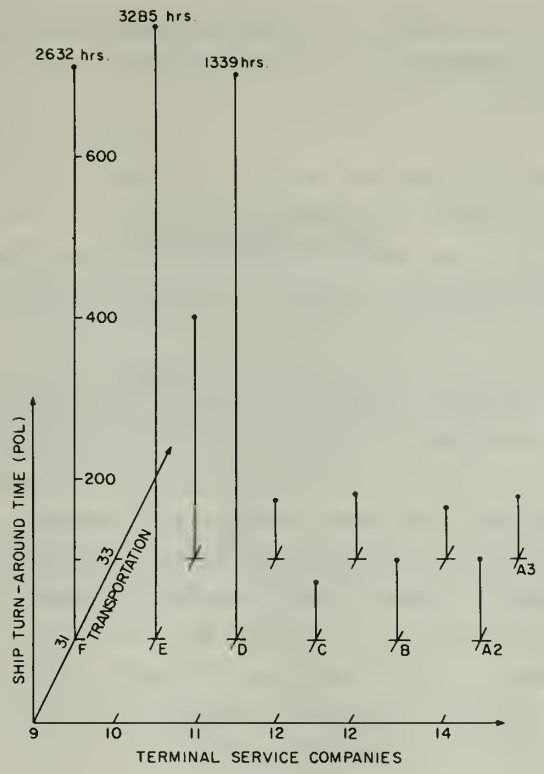



FIGURE 5. Response surface points, POL tanker turn-around.

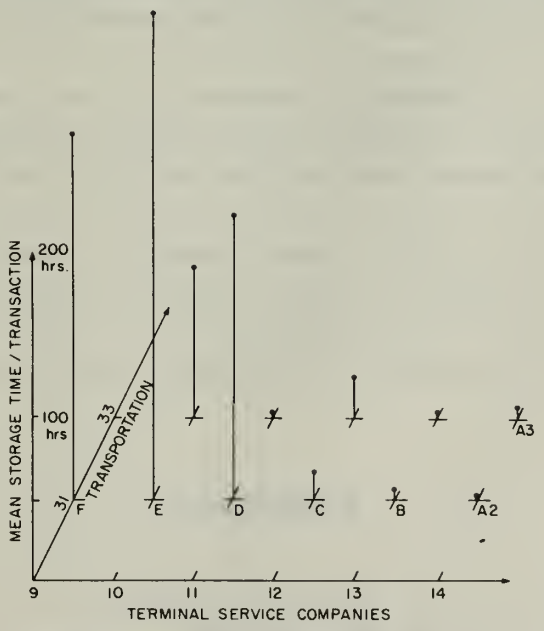



Figure 6. Response surface points, temporary storage.

choose a terminal service level on this basis. Of course, some troop ceiling or constraint is needed to give this method of selection real significance. It is also evident that the gradient is nearly flat until it begins to ascend, and that it then increases markedly. This would indicate that the logistics planner

must pay close attention to the accuracy of arrival input data, to include any anticipated changes, especially in an increasing direction.

## CONCLUSIONS

The difficulties of quantitative analysis were due to the nature of the problem under study; however, a worthwhile qualitative analysis was readily afforded from the simulation's results. Two specific conclusions of significant value were derived. First, the simulation clearly shows that the standard procedures for determining discharge and clearance capacities take insufficient account of the effects of variability in the system. The smallest satisfactory troop level was provided in operating policy D–3. This policy indicated a 253 percent commitment of transportation and a 138 percent commitment of terminal service units compared to the levels determined by prescribed methods. Second, the response surfaces for turn-around times and temporary storage levels indicate that an extremely steep gradient exists as a function of troop levels. Therefore, arrival rates must be accurately predicted, or a safety margin must be provided when troop ceilings are determined.

The approach to the problem in this paper has been at the field manual level, using the kind of data and information available to a general staff. The effort has been to make this model as general and also as usable as the field manual. The tradeoff between efforts in adapting this simulation to a real problem and in preparing a detailed, highly sophisticated simulation of a specific harbor complex is particularly significant. It can be measured in man-hours versus man-years and thus presents a planning aid for rapid identification of feasible alternatives. In practice, it appears that such alternatives are dependent on planning procedures contained in current field manuals. Although the conclusions of this paper simply verify what has been well established in the analytical quarters, they do demonstrate a significant departure from the existing field manual planning values. This demonstrated departure represents the real contribution of the paper.

To be sure such factors as variability in ship arrivals, ship loads, berth productivity, inland transport, and types of commodities are important considerations in any particular port problem; however, there is always the tradeoff between model complexity and ease of modeling. GPSS III is capable of incorporating such complexities, but since current doctrinal references do not yet reflect the dictates of those established analytical principles, it seems appropriate to sacrifice complexity for simplicity in an effort to illustrate the fallibility of currently prescribed procedures for determining discharge and clearance capacities. This model is an effort at such an illustration using an "off the shelf" capability. It is, hopefully, understandable at the field manual level and adaptable to general staff problems and procedures.

## REFERENCES

[1] Besson, F. S. (General, U.S. Army), "From Factory to Foxhole—A 10,000 Mile Pipeline to War," *U.S. News & WorldReport* (19 June 1967), pp. 98–99.

[2] Garvin, Crump (Major General, U.S. Army, Ret.), "Pitfalls in Logistic Planning," *Military Review*, Vol. XLII, No. 4 (April 1962).

[3] FM101–10–1, U.S. Army, *Staff Officer's Field Manual; Organization, Technical and Logistical Data* (January 1966).

[4] FM55–15, U.S. Army, *Transportation Reference Data* (Initial Draft) (March 1967), Chapters 3–5.

[5] IBM, *General Purpose Systems Simulator III* (1965).

[6] ———, *General Purpose Systems Simulator III User's Manual* (1966).

[7] McMillan, Claude and Richard F. Gonzales, *Systems Analysis, A Computer Approach to Decision Models* (Irwin, Inc., 1965), pp. 180–183.

## BIBLIOGRAPHY

Nayler, T. H., et al., *Computer Simulation Techniques* (John Wiley and Sons, Inc., New York, 1966), pp. 248–278.

RB 101–2, USACGSC, *Tables of Organization and Equipment* (April 1964).

RB 101–3, USACGSC, *Combat Service Support* (June 1964).

* * *

# A NOTE ON A MODIFIED PRIMAL-DUAL ALGORITHM TO SPEED CONVERGENCE IN SOLVING LINEAR PROGRAMS

Harold Greenberg

*Naval Postgraduate School*
*Monterey, California*

**ABSTRACT**

The primal-dual algorithm is modified in a two part procedure. In the first part, the pivot row is selected so that an artificial variable is always dropped. The end of the first part usually produces some basic variables with negative values. The second part consists of selecting the most negative basic variable. The equation, represented by the selected basic variable, is multiplied through by minus one and then added to all equations with negative basic variables; it is then augmented by an artificial variable. This procedure produces feasibility for all basic variables and maintains canonical form. The standard primal-dual method is then used to complete the solution. Computational results are presented.

## I. INTRODUCTION

The primal-dual algorithm seems [3] to be theoretically superior to the two-phase simplex method for solving linear programs.

In this paper, we present a modified primal-dual algorithm that achieves faster convergence than the standard primal-dual algorithm. The method also appears to be superior to the two-phase method. Computational results are given.

## II. MODIFICATION OF THE PRIMAL-DUAL ALGORITHM

We consider first the primal-dual algorithm as it appears in Dantzig [2]. By use of the matrix notation, the linear programming problem is: find $x=(x_1, x_2, \ldots, x_n)$ that minimizes

$$z=z_0+cx,$$

when

$$Ax=a \tag{1}$$
$$x_j \geq 0, \qquad j=1, \ldots, n,$$

where $z_0$ is a constant,

c is a 1 by $n$ vector with components $c_j$, $j=1, \ldots, n$,

$a$ is an $m$ by 1 vector with components $a_i \geq 0$, $i=1, \ldots, m$, and

$A$ is an $m$ by $n$ matrix with elements $a_{ij}$, $i=1, \ldots, m$; $j=1, \ldots, n$.

The standard primal-dual algorithm consists of augmenting the constraint equations with artificial variables such as

$$Ax+y=a, \tag{2}$$

where $y$ is an $m$ by 1 vector with components consisting of the artificial variables $y_i$, $i=1, \ldots, m$, and forming

$$w=w_0+dx,$$

where

$$w_0 = \sum_{i=1}^{m} a_i,$$

$d$ is a 1 by $n$ vector with components $d_j, j = 1, \ldots, n$, and

$$d_j = -\sum_{i=1}^{m} a_{ij}.$$

We assume that the components of $c$ have been adjusted before the augmentation by artificial variables so that a feasible solution of the dual problem is readily available. This adjustment is accomplished as in Beale [1]. Thus, we can assume that $c_j \geq 0$, for $j = 1, \ldots, n$.

The problem is to find $x_j \geq 0$, $w = 0$, that minimizes z subject to (2). At any stage of the iteration procedure the equations have the form:

$$\begin{aligned} x_B + \bar{B}x &= \bar{b} \\ y + \bar{A}x &= \bar{a} \\ \bar{d}x &= w - \bar{w}_0 \\ \bar{c}x &= z - \bar{z}_0, \end{aligned} \tag{3}$$

where $x_B$ is a vector representing the current basic variables, $x$ is a vector representing the current non-basic variables, $\bar{B}$ is a matrix with elements $\bar{b}_{ij}$, and $\bar{b}$ is a vector with components $\bar{b}_i$. The matrices and vectors with bars indicate the results of the usual pivot operations within the primal and dual problems. When an artificial variable becomes non-basic, it is dropped from further consideration. This results in the $B$ matrix and the $\bar{b}$ vector in (3). Each iteration in the standard primal-dual algorithm consists of the following:

1. Finding the most negative $\bar{d}_j$ value for those indices $j$ having $\bar{c}_j = 0$.

2. If there are no indices with this property, the $w$ equation is used to change the $z$ equation and force at least one $\bar{c}_j$ to equal zero.

3. When the minimum $\bar{d}_j$ is found with index $j = r$, $x_r$ is selected to be basic.

4. Then min $(\bar{b}_i/\bar{b}_{ir}, \bar{a}_k/\bar{a}_{kr})$ is found for $\bar{b}_{ir} > 0$ and $\bar{a}_{kr} > 0$, where the $i$ and $k$ indices are over the $x_B$ or the $y$ set of equations, respectively. This will select a pivot row in either the $x_B$ set of equations or the $y$ set of equations.

5. The simplex pivot method is then used. If the pivot row is in the $y$ set of equations, the artificial variable in that row is dropped. Thus, in any case, one of the forms in (2) is achieved.

When $\bar{w}' = 0$, the basic solution is optimal and the problem is solved.

The method should include rules to prevent cycling, such as the perturbation technique or the use of lexicographic ordering.

The modification of the above consists of two parts. In the first part 4.) is changed to 4.′). Then min $(\bar{a}_k/\bar{a}_{kr})$ is found for $\bar{a}_{kr} > 0$. This will select a pivot row in the $y$ set of equations. When $\bar{w}_0 = 0$ the basic solution is optimal and the problem is solved if all $\bar{b}_i \geq 0$. Otherwise, we begin the second part of the modification, which consists of:

1. Selecting the equation with most negative $\bar{b}_i$.

2. Multiplying the selected equation through by minus one.

3. Adding the resultant equation to all equations with negative $\bar{b}_i$.

This procedure produces feasibility for all basic variables and maintains canonical form except for the equation developed in step 2, which is then augmented by an artificial variable. We form the $w$

equation in the usual way and then use the standard primal-dual method to complete the solution. The solution so achieved must be optimal since we produce a basic feasible solution while maintaining all $\bar{c}_j \geq 0$.

## III. COMPUTATIONAL RESULTS

Several problems were run on an IBM 360/67 using both the standard and modified primal-dual algorithms. The problems used were test problems from the SHARE Linear Programming Project. Results are given in Table 1.

TABLE 1. *Table of Comparative Results*

| Problem | Size ($m \times n$) | Simplex iterations | Modified primal-dual | | Standard primal-dual | |
|---|---|---|---|---|---|---|
| | | | Iterations | Seconds | Iterations | Seconds |
| Share 1A | $33 \times 64$ | 70 | 39 | 7 | 39 | 7 |
| Share 1D | $27 \times 45$ | 61 | 38 | ..... | 38 | ..... |
| Share 1E | $31 \times 106$ | 66 | 96 | 30 | 136 | 42 |
| Share 1F | $66 \times 135$ | 184 | 125 | 86 | 173 | 114 |
| Share 2A | $30 \times 103$ | 114 | 79 | 29 | 119 | 43 |

The column for simplex iterations was taken from [4] which contains the results for the two-phase method. In these test problems (the only ones attempted), the modified primal-dual is superior to the standard procedure. In problem 1E only did the two-phase method produce less iterations than the modified primal-dual algorithm. On the basis of these test problems the modified primal-dual algorithm is generally superior to the two-phase method as well as the standard primal-dual method.

The author wishes to thank Leola Cutler who kindly supplied the test problems and John E. Easterbrook who helped in the computation and programming.

## REFERENCES

[1] Beale, E. M. L., "An Alternative Method for Linear Programming," Proc. Cambridge Phil. Soc., Vol. 50, No. 4 (1954).

[2] Dantzig, G. B., *Linear Programming and Extensions* (Princeton University Press, 1963).

[3] Hadley, G., *Linear Programming* (Addison-Wesley Publishing Company, Inc., 1962).

[4] Wolfe P. and L. Cutler, "Experiments in Linear Programming," *Recent Advances in Mathematical Programming* (R. L. Graves and P. Wolfe. Eds.) (McGraw-Hill Book Company, Inc., 1963).

* * *

## INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Washington, D.C. 20360. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 50 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

JUNE 1969
VOL. 16, NO. 2
NAVSO P-1278

CONTENTS



OFFICE OF NAVAL RESEARCH
Washington, D.C. 20360